8-2017

Bayesian approach on short time-course data of protein phosphorylation, casual inference for ordinal outcome and causal analysis of dietary and physical activity in T2DM using NHANES data.

You Wu
University of Louisville

BAYESIAN APPROACH ON SHORT TIME-COURSE DATA OF PROTEIN PHOSPHORYLATION, CAUSAL INFERENCE FOR ORDINAL OUTCOME AND CAUSAL ANALYSIS OF DIETARY AND PHYSICAL ACTIVITY IN T2DM USING NHANES DATA

By

You Wu
B.S., Hebei University, 2011
M.S., University of Cincinnati, 2013

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2017

BAYESIAN APPROACH ON SHORT TIME-COURSE DATA OF PROTEIN PHOSPHORYLATION, CAUSAL INFERENCE FOR ORDINAL OUTCOME AND CAUSAL ANALYSIS OF DIETARY AND PHYSICAL ACTIVITY IN T2DM USING NHANES DATA

By

You Wu
B.S., Hebei University, 2011
M.S., University of Cincinnati, 2013

A Dissertation Approved on

August 7, 2017

by the following Dissertation Committee:

Maiying Kong, Ph.D., Dissertation Director

Susmita Datta, Ph.D., Dissertation Co-director

Jeremy Gaskins, Ph.D.

Qi Zheng, Ph.D.

Bert Little, Ph.D.

DEDICATION

This dissertation is dedicated to my parents

Mr. Yuejun Wu

and

Mrs. Linan Chen

whose unconditional love and supports have encouraged me to move ahead.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisors Dr. Maiying Kong and Dr. Susmita Datta for their insightful guidance and constant inspiration over the last four years. I also always deeply appreciate their encouragement when I feel pressured, their sincere advice when I encounter difficulties, and their excellent directions when I get lost. This dissertation and my research would not have been possible without their support and guidance.

I would like to thank Drs. Bert Little, Jeremy Gaskins and Qi Zheng for their time to serve in my dissertation committee. I especially would like to give my thanks to Dr. Gaskins for his great assistance in the pTyr project, and to Dr. Little for his insightful guidance in the causal analysis for dietary and physical activity in T2DM.

I would like to thank the Department of Bioinformatics and Biostatistics for providing me a great environment to do research, and all the faculties and students in the department who have helped me directly and indirectly to complete this dissertation.

Finally, I would like to express my thanks to my parents Mr. Yuejun Wu and Mrs. Linan Chen and my aunt Mrs. Meijuan Chen for their immeasurable supports in my life, so that I can grow to be a person with a strong heart to face difficulties. I am also grateful to my friends: Ruilong Zhang and Fan Zhang, who have given me great encouragement and stood by me. I always feel lucky and grateful to have them in my life.

ABSTRACT

BAYESIAN APPROACH ON SHORT TIME-COURSE DATA OF PROTEIN PHOSPHORYLATION, CAUSAL INFERENCE FOR ORDINAL OUTCOME AND CAUSAL ANALYSIS OF DIETARY AND PHYSICAL ACTIVITY IN T2DM USING NHANES DATA

You Wu

August 7, 2017

This dissertation contains three different projects in proteomics and causal inferences. In the first project, I apply a Bayesian hierarchical model to assess the stability of phosphorylated proteins under short-time cold ischemia. This study provides inference on the stability of these phosphorylated proteins, which is valuable when using these proteins as biomarkers for a disease. in the second project, I perform a comparative study of different confounding-adjusted to estimate the treatment effect when the outcome variable is ordinal using observational data. The adjusted U-statistics method is compared with other methods such as ordinal logistic regression, propensity score based stratification and matching. In the third project, I perform a causal analysis of the combination of dietary information and physical activity in type 2 diabetes across different ethnic groups: White, African American and Mexican American. Such information may contribute to a better understanding of type 2 diabetes variation between ethnic groups, and a better understanding of type 2 diabetes among different ethnic groups and between female and male.

TABLE OF CONTENTS

PAGE

DEDICATION iii

ACKNOWLEDGMENTS iv

ABSTRACT v

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1: INTRODUCTION 1
1.1 Bayesian Approach on Protein Phosphorylation Study 1
1.2 Comparative Study for Ordinal Outcomes using Observational Data . 2
1.3 Causal Analysis of T2DM using Path Analysis 3

CHAPTER 2: PROFILING THE EFFECTS OF SHORT TIME-COURSE COLD ISCHEMIA ON TUMOR PROTEIN PHOSPHORYLATION USING A BAYESIAN APPROACH 3
2.1 Introduction 4
2.2 Methods 6
2.2.1 Hierarchical Bayesian model 6
2.2.2 Estimation and inference of profile classification 9
2.2.3 ORI method 11
2.3 Application to an ovarian tumor dataset 13
2.3.1 The ovarian tumor data 13
2.3.2 Results 14
2.4 Simulation Study 16
2.4.1 Assessing the effect of sample size 17
2.4.2 Assessing the effect of the signal-to-noise ratio and number of time points 18
2.5 Discussion and Conclusion 20
2.6 Tables and Figures 22

CHAPTER 3: INVESTIGATION OF DIFFERENT STATISTICAL METHODS FOR ESTIMATING TREATMENT EFFECTS WHEN OUTCOME VARIABLE IS ORDINAL AND CONFOUNDING EXISTS 26
3.1 Introduction 30
3.2 Statistical Method 33
3.2.1 Parametric method for estimating superiority score 34

3.2.2 Propensity score based methods for estimating superiority score 35
3.2.3 Adjusted Mann-Whitney U test statistic 38
3.3 Simulation Study 40
3.3.1 Scenario 1: Outcome follows an ordinal logistic model 41
3.3.2 Scenario 2: Outcome follows a mixture cumulative Box-Cox model 42
3.4 NHANES data application 44
3.5 Discussion and Conclusion 45
3.6 Tables and Figures 47

CHAPTER 4: CAUSAL ANALYSIS OF DIETARY INFORMATION AND PHYSICAL ACTIVITY IN TYPE 2 DIABETES BY GENDER IN WHITE, AFRICAN AMERICAN AND MEXICAN AMERICAN: NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEYS 2011-2014 49
4.1 Introduction 51
4.2 Method 53
4.2.1 Data and Materials 53
4.2.2 Statistical Analysis 55
4.3 Results 56
4.4 Discussion and Conclusion 59
4.5 Tables and Figures 64

REFERENCES 80

APPENDIX 92

CURRICULUM VITA 99

LIST OF TABLES

TABLE PAGE

2.1 Detailed Result of Bayesian classification based on MAP. The probability of each protein being classified to the null profile is reported in 4^{th} column with the column title "$\hat{P}(Z_j = 0|Y)$". The posterior probabilities of classifying each protein to its corresponding profile is reported in the 5^{th} column with the column title "Maximum Posterior Probabilities". The candidate profiles included in 75% credible set for each protein are reported in the last column. * indicates that the pTyr site is classified to the same profile using both the ORI method and the Bayesian model. 23

2.2 Posterior median and 95% CI for each model parameter 24

2.3 Comparison of classifications from the ORI and MAP estimators for the human tumor data. 24

2.4 Simulation results from Section 2.4.1. 200 simulated datasets with $J = 35$ and $T = 4$ were used in both the simulation settings. In each simulation, 8 pTyr sites were assigned to the null profile C_0 and 27 pTyr sites were assigned to the other candidate profiles. The quantities reported in this table are averaged across 200 simulated datasets. The accuracy of the estimated trajectory is measured by the sum of squared errors (SSE), and a lower value indicates a higher accuracy. 24

2.5 Simulation results from section 2.4.2. The correct classification rates is formed under different T and different effect sizes. For each simulation study, $N = 20$, $J = 35$ and 200 simulated datasets were used. The accuracy of the estimated trajectory is measured by the sum of squared errors. 25

3.1 Sizes of the tests based on different methods for simulated data sets generated under simulation Scenario 1. 47

3.2 Summarized simulation results for Scenario 1, where outcome was generated from an ordinal logistic regression model. 48

3.3 Summarized simulation results for Scenario 2, where the outcome variable was generated from a mixture of Box-Cox distributions. 48

4.1 Descriptive statistics for each variable in the analysis. Mean (SD) is reported for the continuous variable by gender and race. N (%) is reported for categorical variable. 65

4.2 Significant paths and summarized causal effects for each variable in the path analysis for the entire population. Dietary variables in the path analysis include protein, total sugars, total fat, magnesium, moisture and butanoic fatty acid. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$". 66
4.3 Significant paths and summarized causal effects for each variable in the path analysis for male. Dietary variables in the path analysis include energy, protein, carbohydrate, total sugars, total fat and moisture. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$". 67
4.4 Significant paths and summarized causal effects for each variable in the path analysis for female. Dietary variables in the path analysis include energy, protein, carbohydrate, total sugars, total fat and moisture. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$". 68
4.5 Significant paths and summarized causal effects for each variable in the path analysis for Mexican American by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. 68
4.6 Significant paths and summarized causal effects for each variable in the path analysis for non-Hispanic black by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. 69
4.7 Significant paths and summarized causal effects for each variable in the path analysis for non-Hispanic white by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. 70
A1.1 All possible candidate profiles with nodal parameters for $T = 4$. Nodal parameters for ORI method are defined as the parameters linked to all the other parameters where the two parameters are considered as linked if the inequality between them is pre-specified. 92
A1.2 Detailed Results of ORI method for ovarian tumor study. P-values for test against null are reported in the last column. * indicates the pTyr site is classified to the same profile using both ORI method and the MAP estimator of Bayesian model. 93
A1.3 Selected profiles and the percentage of $J = 35$ sites assigned to each profile for $T = 5$ and $T = 8$ in simulation study of Section 2.4.2. . . . 94

LIST OF FIGURES

FIGURE PAGE

2.1 Estimated trajectory with 95% credible interval (1^{st} and 3^{rd} columns) and MPW estimators (2^{nd} and 4^{th} columns) for six representative phosphorylated proteins. The gray lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications. 26

2.2 Boxplots of simulation results from Section 2.4.2. The correct classification proportions shown are calculated based on 200 simulated datasets for each case. For $T = 4$, all the possible candidate profiles are considered. For $T = 5$ and 8, the ORI method is not used when considering all possible profiles due to the limitation of computation speed. 27

2.3 Boxplots of sum of squared error (SSE) measuring the accuracy of trajectory estimation. A lower value indicates higher accuracy. 28

2.4 Q-Q plot and histogram of estimated residuals for ovarian tumor data. 29

3.1 Power curves for different methods with data generated from ordinal logistic regression models in simulation Scenario 1. 49

3.2 Power curves for different methods in with data generated under simulation Scenario 2. 50

4.1 Hypothesis causal model (Adapted from Wright, 1934). X_1 and X_2 indicate demographic characteristics or physical activity. N_1 and N_2 are nutrition intake variables, which reflect the dietary information. . 71

4.2 Path diagram for the entire sample. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 72

4.3 Path diagram for male. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients,where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 73

4.4 Path diagram for female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 74

4.5 Path diagrams for Mexican American. Left panel is the diagram for Mexican American male and right one is for Mexican American female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 75
4.6 Path diagrams for non-Hispnic black female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 76
4.7 Path diagrams for non-Hispnic white male. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. . 77
4.8 Path diagrams for non-Hispnic white female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients,where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 78
4.9 Path diagram for whole population using two dummy variables to measure BMI. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram. 79
A1.1 Estimated Trajectory with 95% credible interval ($1st$ and $3rd$ columns) and MPW estimators ($2nd$ and $4th$ columns) of proteins 1-10. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications. 95
A1.2 Estimated Trajectory with 95% credible interval ($1st$ and $3rd$ columns) and MPW estimators ($2nd$ and $4th$ columns) of proteins 11-20. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications. 96

A1.3 Estimated Trajectory with 95% credible interval ($1st$ and $3rd$ columns) and MPW estimators ($2nd$ and $4th$ columns) of proteins 21-30. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications. 97
A1.4 Estimated Trajectory with 95% credible interval ($1st$ and $3rd$ columns) and MPW estimators ($2nd$ and $4th$ columns) of proteins 31-32. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications. 98

CHAPTER 1

INTRODUCTION

1.1 Bayesian Approach on Protein Phosphorylation Study

Phosphorylated proteins provide insight into tumor etiology and are used as diagnostic, prognostic and therapeutic markers of complex diseases. However, pre-analytic variations, such as freezing delay after biopsy acquisition, often occur in real hospital settings and potentially lead to inaccurate results. The objective of the first project is to develop statistical methodology to assess the stability of phosphorylated proteins under short-time cold ischemia. We consider a hierarchical model to determine if phosphorylation abundance of a protein at a particular phosphorylation site remains constant or not during cold ischemia. When phosphorylation levels vary across time, we estimate the direction of the changes in each protein based on the maximum overall posterior probability and on the pairwise posterior probabilities, respectively. We analyze a dataset of ovarian tumor tissues that suffered cold-ischemia shock before the proteomic profiling. Gajadhar et al. (2015) applied independent clusterings for each patient because of the high heterogeneity across patients, while our proposed model shares information allowing conclusions for the entire sample population. Using the proposed model, 15 out of 32 proteins show significant changes during one-hour cold ischemia. Through simulation studies we conclude that our proposed methodology has a higher accuracy for detecting changes compared to an order restricted inference method. Our approach provides inference on the stability of these phosphorylated proteins, which is valuable when using these proteins as biomarkers for a disease.

This work is published The details of the work are presented in Chapter 2.

1.2 Comparative Study for Ordinal Outcomes using Observational Data

Ordinal outcomes are frequently observed in the clinical studies, and also in the social and economic sciences. The commonly used methods for analyzing ordinal outcome include the parametric statistical methods (e.g., ordinal logistic regression model) and the non-parametric statistics (e.g., the Mann-Whitney U test statistic). However, the ordinal logistic regression model may be less robust when the model is misspecified, and the classic Mann-Whitney U test does not have control of the confounding covariates which may result in seriously biased estimates in the observational studies. The propensity score based methods, such as matching, stratification, and inverse probability weighting method, have been applied to assess treatment effect with control of confounding covariates. However, these methods usually assign the ordinal outcome variables to numerical scores in the analysis, thus losing information about the nature of ordinal variables. Glen, Kong and Datta (2017) propose an adjusted U-statistic to estimate the treatment with control of the confounding covariates by using the inverse probability weighting, which weights each subject into its study population, and the weight for a subject is obtained from the propensity score. In this second project, we provide a comparative study of different methods for estimating treatment effects for ordinal outcome variables using observational data. We compare the adjusted U-statistics method (i.e. the weighted Mann-Whitney U statistics) with other popular used methods such as ordinal logistic regression, propensity score based stratification and matching. Extensive simulation studies are carried out to compare the performance of these methods under different situations, and pros and cons are presented. A case study is constructed to assess the effect of physical activity on diabetic status

with controlling the confounding of the sociodemographic characteristics and dietary information. This project is presented in Chapter 3.

1.3 Causal Analysis of T2DM using Path Analysis

Type 2 diabetes mellitus (T2DM) has become a major public health problem worldwide and one of the major causes of mortality in the United States outstripping cancer, HIV/AIDS, and cardiovascular disease. Enormous economic and psychosocial consequences are also associated with T2DM (Menke et al., 2015; Zimmet et al., 2016; Wang et al., 2016). It is important to examine the relationship between diet and type 2 diabetes because food intake is considered as the crucial variable in the control of T2DM. The second important component of T2DM control is physical activity. The National Health and Nutrition Examination Survey (NHANES) continuous database includes physical activity and detailed information on dietary components. These data have not been previously analyzed and may offer important insights into dietary variability, physical activity and the management of T2DM. The objective of the third project in this dissertation is to analyze the causal paths that predict type 2 diabetes using data on demographics, dietary, BMI and physical activity. The analytical model is path analysis (i.e., causal structural equations) to explore the relationships between variables that are clearly causal and outcomes (i.e., type 2 diabetes status). Via the causal analysis, the causal importance of these variables on type 2 diabetes can be quantitatively examined. This work is presented in Chapter 4.

CHAPTER 2

PROFILING THE EFFECTS OF SHORT TIME-COURSE COLD ISCHEMIA ON TUMOR PROTEIN PHOSPHORYLATION USING A BAYESIAN APPROACH[1]

2.1 Introduction

Protein tyrosine phosphorylation is considered to be a fundamental mechanism for regulating many cellular functions. In this specific phosphorylation, a phosphate group is added to the amino acid tyrosine on a protein. Disorders of tyrosine phosphorylation are believed to lead to many serious human diseases (Hunter, 2009). For example, protein tyrosine phosphorylation is tightly regulated in normal cells, but tyrosine kinases, whose activity controls tyrosine phosphorylation, are found to be mutated or over-expressed in many human malignancies (Paul and Mukhopadhyay, 2004). Accurate and robust assessment of tyrosine phosphorylation in tumor biopsy samples is thus necessary for understanding intracellular signaling networks and for developing targeted therapies for cancer patients (Bonnas et al., 2012; Gajadhar et al., 2015).

However, there may exist pre-analytic variations due to inconsistencies during sample collection and processing in a clinical laboratory. One of these sources of variations is cold ischemia. Also known as freezing delay time, cold ischemia is the

[1]Reproduce with permission from "Profiling the effects of short time-course cold ischemia on tumor protein phosphorylation using a Bayesian approach" by You Wu, Jeremy Gaskins, Maiying Kong and Susmita Datta, 2017. Biometrics. doi:10.1111/biom.12742. Copyright © 2017, The International Biometric Society.

time between tissue specimen excision and the freezing of the sample. Although it has been shown that global protein levels do not change up to one hour cold ischemia, significant changes are observed in phosphorylated proteins at the phosphorylation sites. Some of the phosphorylation sites even have rapid changes during the first 5 minutes of cold ischemia (Mertins et al., 2014; Gajadhar et al., 2015). The dynamic nature of protein phosphorylation in tissue specimens may be due to the fact that the kinases and phosphatases controlling the signaling pathways of phosphorylation are still active *ex vivo* after tissue excision (Espina et al., 2008). Therefore, the fidelity of a protein phosphorylation abundance, which is the relative intensity value used to measure protein phosphorylation magnitude, cannot be guaranteed in excised tumor tissues that undergo cold ischemia. Subsequently, the targeted therapeutic strategies and clinical decisions based on phosphorylation studies may not be accurate. To ensure trustworthy results, it is necessary to examine whether the phosphorylation level of a specific protein at a specific site is affected by short time cold ischemia. This may determine the suitability of that phosphorylated protein to be considered as a stable biomarker for a disease.

Although some phosphorylated proteins have been observed to be affected by short time cold ischemia in the literature, few of those studies comprehensively examine the stability of phosphorylation over an entire sample population (Gajadhar et al., 2015; Espina et al., 2008; Gündisch et al., 2013). As mentioned by Montana, Berk and Ebbels (2011), most metabolomic experiments only produce short time-course data with less than 10 time points, and the classical time series analysis cannot be applied. Gajadhar et al. (2015) examined the cold ischemia induced changes in phosphorylation by applying an affinity propagation clustering analysis to an ovarian tumor dataset from 5 patients. However, due to the high level of heterogeneity across patients, an independent clustering was formed for each patient, and it is challenging to draw conclusions for the entire population. In this article, we construct a novel

hierarchical Bayesian model to examine whether the phosphorylation of each protein at a particular site stays stable or not under cold ischemia shock. When the phosphorylation levels at a particular site vary across time points, we further estimate the direction of the changes based on whether their abundances are increasing or decreasing between two adjacent time points. By utilizing random effects to capture the dependence between the observed phosphorylation abundances across different time points and different phosphorylated proteins, we develop a model that shares information across patients and allows us to draw conclusions for the entire sample population.

Our proposed Bayesian model is presented in Section 2.2. We also briefly describe an existing method proposed by Peddada et al. (2003) which can potentially be used to analyze this short time course data. The work of Peddada et al. (2003) is based on order restricted inference (ORI) for the analysis of a short time-course microarray data. In Section 2.3, we apply our proposed method and the competitor ORI approach to examine the stability of each phosphorylated protein under one hour cold ischemia shock using an ovarian tumor dataset from Gajadhar et al. (2015). We further verify our method using a simulation study in Section 2.4 and provide some concluding remarks in Section 2.5.

2.2 Methods

2.2.1 Hierarchical Bayesian model

Suppose that there are N patients in the study. We consider J common phosphorylated proteins, and each protein has only one phosphotyrosine (pTyr) site. For each patient the abundance of phosphorylation at the j^{th} pTyr site ($j = 1, \ldots, J$) is measured at T common time points. At the t^{th} time point ($t = 1, \ldots, T$), A_t represents the number of minutes of freezing delay after sample excision (i.e., minutes of cold

ischemia). The abundance of phosphorylation at the j^{th} pTyr site at the t^{th} time point for the i^{th} patient is denoted as Y_{ijt} $(i = 1, \ldots, N;\ j = 1, \ldots, J;\ t = 1, \ldots, T)$. The unknown true population mean of the response profile at the j^{th} pTyr site is expressed as $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jT})'$.

Our primary goal is to examine the stability of each phosphorylated protein and then further classify the detected unstable proteins according to their changes in phosphorylation abundance between two adjacent time points. We characterize the changes for each site into $c = 2^{T-1} + 1$ classes. The null profile is set as $C_0 = \{\mu_j \in R^T : \mu_{j1} = \mu_{j2} = \cdots = \mu_{jT}\}$, indicating that there is no change in phosphorylation abundance under cold ischemia shock over time. Each of the remaining classes has the form $C_r = \{\mu_j \in R^T : \mu_{j1} \square \mu_{j2} \square \cdots \square \mu_{jT}\}$ $(r = 1, \ldots, 2^{T-1})$, where $\square$ is either $<$ or $>$. Our primary objective is to learn if $j \in C_0$ versus any of the other classifications. The secondary objective is to learn which C_r is the most probable class for sites not in the null profile.

To model the changes of phosphorylation during one hour of cold ischemia, we consider a hierarchical random effect model of the following form:

$$Y_{ijt} = \mu_{jt} + \gamma_i + \eta_{ij} + \delta_{it} + \varepsilon_{ijt}. \tag{2.1}$$

The main quantity of interest μ_{jt} represents the mean abundance of phosphorylation at the j^{th} pTyr site and the t^{th} time point. Because the measurements for each patient across multiple sites and multiple time points are dependent, appropriate techniques are required to make an efficient and valid inference. The overall variation across patients is explained by γ_i $(i = 1, \ldots, N)$. Variation across different pTyr sites within the i^{th} patient is explained by the site-specific random effect $\eta_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iJ})'$. The patient-specific temporal effect is captured by the random effect $\delta_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{iT})'$, representing the time effect for the i^{th} subject. The

random error ε_{ijt} is assumed to be normally distributed with mean 0 and variance σ^2.

In this work, a Bayesian approach is considered for the inference. Due to normality of the random errors, the data is distributed as $Y_{ijt} \sim N(\mu_{jt} + \gamma_i + \eta_{ij} + \delta_{it}, \sigma^2)$. The overall effect from the i^{th} patient, γ_i, is modeled by a normal distribution with mean 0 and variance σ^2_γ. The site-specific random effect within the i^{th} patient is modeled by $\eta_{ij} \sim N(0, \sigma^2_\eta)$ $(j = 1, \ldots, J)$. The distribution of the temporal effects is $\delta_i \sim MVN(\mathbf{0}, \sigma^2_\delta R(\rho))$. We assume an auto-regressive correlation structure for $R(\rho)$ with $corr(\delta_{it}, \delta_{it'}) = \rho^{|A_t - A_{t'}|/s}$, although other choices are possible. Here, ρ represents the correlation between responses s minutes apart, and we use the value of $s = 10$ minutes throughout. Recall that A_t represents the number of minutes of cold ischemia.

Our primary goal is to determine if there are overall changes in the abundance at the j^{th} site, that is, if $j \in C_0$ or if j is in one of the other profiles. To that end, we introduce the random variable $Z_j \sim Bernoulli(\theta)$ to indicate whether there is variation in phosphorylation abundance at the j^{th} pTyr site during cold ischemia. The parameter θ represents the overall proportion of sites with ischemia-induced changes. The prior distribution of the trajectory μ_{jt} is set as $\mu_{jt}|(\mu^*_j, Z_j) \sim N(\mu^*_j, Z_j\sigma^2_\mu)$, where $\mu^*_j \sim N(\mu_0, \sigma^2_{\mu^*})$ gives the average value across all the T time points. If $Z_j = 0$, the μ_{jt}s are equal to μ^*_j across all the time points, i.e., the j^{th} pTyr site is classified to the null profile C_0. If $Z_j = 1$, μ_{jt} varies across time, and the j^{th} site is impacted by cold ischemia. A site with $Z_j = 1$ is classified to one of the candidate profiles $C_1, \ldots, C_{c-1}$ based on the inequality directions in μ_j across all the time points. The prior distribution of the overall mean μ_0 is set as normal distribution with mean 0 and standard deviation 100.

Conjugate prior distributions for the error variance σ^2 and fixed effect variance components σ^2_μ and $\sigma^2_{\mu^*}$ are set as inverse-gamma distribution with both shape and

scale parameters at 0.1. Since we expect that the overall patient effect γ_i, patient-specific temporal effect δ_{it} and site-specific effect η_{ij} have smaller impact on the response compared to μ_j, the prior distributions for the standard deviations σ_γ, σ_η, and σ_δ are set as half-Cauchy distributions with medians (equivalent to the scale parameter) of 0.1. The half-Cauchy is a common prior for the standard deviation of a regression coefficient or shrinkage effect as it has a non-zero density at 0 (unlike inverse-gamma distribution) allowing a small value for the variance term. Further, the heavy tail of the half-Cauchy prior protects from over-shrinkage by allowing large values when the effect is highly influential (Gelman, 2006; Carvalho, Polson and Scott, 2010). The correlation parameter ρ is given a uniform distribution from 0 to 1 as a prior distribution. The prior for θ, the population proportion of sites impacted by cold ischemia, is taken to be $Beta(1, 1)$ (i.e., Uniform[0, 1]).

2.2.2 Estimation and inference of profile classification

Markov chain Monte Carlo (MCMC) sampling is applied to obtain a posterior sample of size M. For each phosphorylation site j, denote the m^{th} sample of the posterior trajectory μ_j as $\mu_j^{(m)}$ $(m = 1, \ldots, M)$. First, we consider testing $H_{0j} : Z_j = 0$ vs. $H_{1j} : Z_j = 1$ to examine whether the j^{th} site is impacted by cold ischemia. $P(Z_j = 1|Y)$ is estimated by the proportion of MCMC iterations with $Z_j^{(m)} = 1$. The j^{th} site is classified to the alternative profiles if $\hat{P}(Z_j = 1|Y)$ is greater than 0.5, which is the Bayes decision rule corresponding to the $0 - 1$ loss function $L(Z_j, \hat{Z}_j) = I(Z_j \neq \hat{Z}_j)$. If we choose to penalize Type I and Type II errors differently, the 0.5 threshold can be adjusted. Placing a non-degenerate prior on θ, the overall probability of changes, guarantees an automatic multiplicity correction over our tests based on Z_j (Scott and Berger, 2010). This approach allows us to maintain the false discovery rate when we consider whether μ_j varies across T time points (i.e., $Z_j = 1$) over a potentially large number of sites J.

If $H_{0j} : Z_j = 0$ is rejected, our secondary objective is to estimate the changes in direction of the unstable phosphorylated proteins. To estimate the profile of the j^{th} site, we consider two competing estimators: the maximum a posteriori (MAP) estimator and maximum pairwise (MPW) estimator. Both estimators are formed conditionally on the conclusions for H_{0j} vs. H_{1j}. If the posterior probability $P(Z_j = 0|Y)$ is larger than the threshold value (typically, 0.5), then the j^{th} site maintains a membership in the null profile C_0, indicating no change in phosphorylation abundance during cold ischemia. Otherwise, we reject H_{0j}, conclude that the j^{th} pTyr site varies across T time points, and classify the j^{th} site to its most likely profile. Using the MAP estimator, we place the j^{th} pTyr site that rejects H_{0j} to the profile that maximizes the estimated posterior probability $\hat{P}(\mu_j \in C_r|Y)$ $(r = 1, \ldots, c-1)$, which is the profile with the majority votes among the M MCMC iterations, $\arg\max_r \sum_m I(\mu_j^{(m)} \in C_r)$.

For the MPW estimator, the pTyr site that rejects H_{0j} is classified to one of the candidate profiles $(C_1, \ldots, C_{c-1})$ by estimating the direction of change between adjacent measurements pairwise. Choosing the inequality between a pair of values $\mu_{j,t-1}$ and μ_{jt} is based on choosing the larger of $\hat{P}(\mu_{j,t-1} < \mu_{jt}|Y)$ and $\hat{P}(\mu_{j,t-1} > \mu_{jt}|Y)$. The sequence of pairwise-estimated inequalities is used to determine the class of the phosphorylated protein.

In general, one would expect that the MAP estimator produces more accurate classification as it takes the full vector μ_j into consideration. However, the MAP estimator is based on the number of iterations that μ_j visits each of the 2^{T-1} non-null profiles C_r during the MCMC run. When T is large relative to the length of the MCMC chain, even the most probable profiles are visited infrequently, and the estimated posterior probabilities $\hat{P}(\mu_j \in C_r|Y)$ may have large variabilities relative to the differences between the more likely profiles. Conversely, the marginal pairwise probabilities $\hat{P}(\mu_{j,t-1} < \mu_{jt}|Y)$ and $\hat{P}(\mu_{j,t-1} > \mu_{jt}|Y)$ typically have small variabilities. Therefore, the MPW estimator may provide preferable classification for a large

T. This issue is similar in spirit to choosing between the MAP and median model estimators in Bayesian variable selection (see e.g., George and McCulloch, 1993; Barbieri and Berger, 2004).

To describe the confidence in classification, we form a credible set $C_{j,1-\alpha}$ for each site j where $\hat{P}(\mu_j \in C_{j,1-\alpha}|Y) \geq 1-\alpha$. For the MAP estimator, the profiles that are included in the $100(1-\alpha)\%$ credible set are determined by their estimated posterior probabilities. For each phosphorylated protein, we sort $\hat{P}(\mu_j \in C_0|Y)$, $\hat{P}(\mu_j \in C_1|Y)$, ..., $\hat{P}(\mu_j \in C_{c-1}|Y)$ in decreasing order. Then the top candidate profiles are included in the credible set until their cumulative posterior probability is greater than $100(1-\alpha)\%$. For the MPW estimator, a similar procedure can be performed using the product of the posterior probabilities between each of the two adjacent time points to approximate the posterior probability of each profile.

2.2.3 ORI method

As a comparison to our proposed methodology, we also apply an order restricted inference (ORI) algorithm developed by Peddada et al. (2003). This ORI method is one of the most commonly used methods developed for analyzing short time-course microarray data. It can also be applied to this phosphorylation abundance data. We briefly describe the method here. The sample mean vector of the j^{th} pTyr site is denoted as $\bar{Y}_j = (\bar{Y}_{j1}, \ldots, \bar{Y}_{jT})'$.

First, a collection of the candidate profiles $C_1, \ldots, C_p$ is specified. The collection may or may not be the full set of 2^{T-1} possible non-zero profiles. Second, for each candidate profile, point estimation follows the procedures proposed by Hwang and Peddada (1994), which use an isotonic regression estimator for the general ordering cases. For each profile, the nodal parameter is the key for the estimation procedure. The nodal parameters in each profile are defined as the parameters which are linked to all the other parameters, where the two parameters are linked if the in-

equalities between them are pre-specified. For example, in the case that $T = 4$, nodal parameters are μ_1, μ_2, μ_3 and μ_4 in the candidate profile $C = \{\mu_1 < \mu_2 < \mu_3 < \mu_4\}$; in the candidate profile $C = \{\mu_1 < \mu_2 < \mu_3 > \mu_4\}$, μ_3 is the only nodal parameter (see Table A1.1 in the Appendix for the details when $T = 4$). For the nodal parameters, an ordered sequence of parameters can be formed as $\mu_{j(1)} \leq \mu_{j(2)} \leq \ldots \leq \mu_{j(T)}$, where one may arbitrarily assign an inequality between the two parameters if their relationship is unknown. Based on the ordered sequence, the estimate of the nodal parameter μ_{jt} is found using the following formula:

$$\hat{\mu}_{jt} = \hat{\mu}_{j(s)} = \min_{r \geq s} \max_{q \leq s} \frac{\sum_{k=q}^{r} n_{j(k)} \bar{Y}_{j(k)}}{\sum_{k=q}^{r} n_{j(k)}},$$

where $\bar{Y}_{j(k)}$ is the sample mean based on $n_{j(k)}$ observations. For a non-nodal parameter, say $\mu_{jt'}$, the largest sub-profile with $\mu_{jt'}$ as a nodal parameter needs to be identified, then $\mu_{jt'}$ is estimated by using the formula above for the nodal parameters and the data corresponding to the sub-profile. For the parameters in a profile without any nodal parameters, the largest sub-profile with at least one nodal parameter needs to be identified, and the parameters in the sub-profile can be estimated using the methods for the profiles with nodal parameters. The procedure is repeated until all of the parameters are estimated. Once a parameter is estimated, the estimate of this parameter should be used in the estimation of the other unknown parameters.

Hypothesis testing for $H_0 : \mu_j \in C_0$ vs. $H_1 : \mu_j \in \cup_{k=1}^{p} C_k$ (equivalent to our test of $H_0 : Z_j = 0$ vs. $H_1 : Z_j = 1$) is performed using a bootstrap approach. In each bootstrap step, the maximum difference of two linked parameters is calculated for each candidate profile. Then the largest maximum difference among the p candidate profiles can be obtained for each bootstrap sample. The hypothesis test is performed based on the bootstrap distribution of the largest maximum difference. If H_0 is rejected, the site is classified to the profile with the largest maximum difference. Full

details can be found in Peddada et al. (2003).

2.3 Application to an ovarian tumor dataset

2.3.1 The ovarian tumor data

The ovarian tumor data from Gajadhar et al. (2015) is used to illustrate our proposed methodology. Ovarian tumor tissues were collected from $N = 5$ patients. Tumor specimens were collected across $T = 4$ time points at 0, 5, 30 and 60 minutes of freezing delay after the surgical removal. After protein extraction, digestion and phosphotyrosine peptides enrichment, the peptides are separated using a mass-spectrometry based analysis. The phosphotyrosine peptides and proteins are identified and quantified subsequently.

Only the phosphorylated proteins with one phosphotyrosine position are considered in the study. The values of phosphorylation abundance are relative intensities determined by iTRAQ quantification. For relative quantification of the four time points, peptide samples were chemically labeled with different iTRAQ reagents and then mixed into a single sample. These tags get covalently attached to the amine groups of the N termini of peptides and lysine side chains. The mixed sample then get fragmented in the mass spectrometer and reporter ions at defined masses (including the masses of the pTyr sites) get released. Ratios of these reporter ions to one another are treated as relative abundance, and the intensity of the masses at the initial time point is treated at the reference abundance (Gajadhar et al., 2015). In order to find the general characteristics of the phosphorylated sites, only the phosphorylated proteins that are common to all patients are included in the following analysis. After excluding the proteins with multiple phosphotyrosine positions, $J = 32$ common pTyr sites are used in the analysis.

2.3.2 Results

The sampling procedure for our Bayesian methods is implemented using WinBUGS (Spiegelhalter et al., 1999) and R (R Core Team, 2014). Copies of the WinBUGS and R codes can be found in the online supplementary materials. Running the MCMC sampling procedure for 50,000 iterations with a burn-in of 5000 and thinning rate of 5 takes 3.02 minutes. The hypothesis $H_{0j} : Z_j = 0$ is rejected if $\hat{P}(Z_j = 1|Y) > 0.5$, and both MAP and MPW estimators are used to classify each phosphorylated protein. The testing and classification results based on these two Bayesian estimators are exactly the same.

Figure 2.1 shows the estimated trajectories with 95% credible intervals (CI) (1^{st} and 3^{rd} columns) and MPW estimators (2^{nd} and 4^{th} columns) of six representative phosphorylated proteins. The gray lines represent the observed phosphorylation abundance of each patient. The boxes in the MPW plots represent the posterior probabilities of an equal, decreasing, or increasing trend of phosphorylation abundance between each pair of adjacent time points, and the shaded box indicates the MPW-chosen classification. For example, both phosphorylation sites of DYRK1A and CDK1 are classified to the null profile C_0 with $\hat{P}(Z = 0|Y) = 0.98$ and 0.95, respectively. Sites of FYB, MAPK7, GAB1 and MAPK14 are classified to the non-null profiles.

Figure 2.1 also shows that some phosphorylation sites have similar phosphorylation trajectories across all five patients (DYRK1A and MAPK7), while some other sites (CDK1 and MAPK14) have greater heterogeneities across patients. For example, the flat trajectory seen in phosphorylated protein CDK1 may be partly due to the conflicting effects of the patient-dependent variations. The plots of estimated profiles and MPW estimator for all phosphorylated proteins can be found in Figures A1-A4 in the Appendix.

Detailed classification results from the Bayesian model are shown in Table

2.1 with the probabilities of being classified to the null profile (4^{th} column) and the posterior probabilities of being classified to the corresponding profiles based the MAP estimator (5^{th} column). The profiles included in the 75% credible set are reported in the last column in Table 2.1. Recall the site's trajectory is estimated to be one of the profiles in this set with at least 75% posterior probability. 17 pTyr sites were classified to the null profile, indicating that the phosphorylation abundance does not change during one-hour cold ischemia based on both estimators. The pTyr sites of DOK1, INPP5D and MPZL1 show an increasing trend (C_1), while the sites of CDKL5, EPHA2, FYB and PRKCD show a decreasing trend during one hour cold ischemia (C_8). Some of our classification results are consistent with the clustering results in Gajadhar et al. (2015). For example, ANXA2 and FLNA were classified to the null profile C_0 based on our proposed model, and they were also classified to the cluster with minimal quantitative fluctuations in Gajadhar et al. (2015). MAPK14 showed a rapid increase in phosphorylation within 5 minutes cold ischemia in Gajadhar et al. (2015), and it was classified to the candidate profile C_3 with similar characterization based on our Bayesian model. However, the conclusions in Gajadhar et al. (2015) are based on a single sample, while our conclusions are based on all samples and should be more robust.

Table 2.2 shows the posterior median and 95% credible interval (CI) for each model parameter. The estimated standard deviations for the patient and site-specific components, σ_γ and σ_η , are quite small, indicating the impact of our half-Cauchy shrinkage prior. Conversely, the standard deviation of the temporal effect σ_δ is much larger (roughly 60% of the standard deviation for error), while the correlation is low ($\rho = 0.13$). Thus, variation across T time points is the key driver of the covariance structure, not the variation across patients or sites. By including the within patient temporal effect and site-specific effect, we are able to cohesively distinguish between the roles of protein trajectory μ_{jt} and the patient trajectory δ_{it}, while Gajadhar et al.

(2015) performs patient-specific analysis which potentially confounds the two sources of variation.

Classification was also done using the ORI method. Detailed results of ORI profiling with p-values for the test of $H_0 : \mu_j \in C_0$ versus $H_1 : \mu_j \in \cup_{k=1}^{c-1} C_k$ can be found in Table A1.2 in the Appendix. Table 2.3 compares the classifications from the Bayesian approach to those from the ORI method. Of the 17 pTyr sites classified to the null profile by the Bayesian method, the ORI method classifies 14 of these to the null profile and 3 to the non-null profiles. There is one profile which the ORI method classifies to the null profile, however, the Bayesian method classifies to the non-null profile. In total, 23 pTyr sites (71%) were classified to the same profiles using both methods. The total computation time using the ORI method was 26.4 minutes, almost seven times longer than our Bayesian method.

2.4 Simulation Study

To verify our approach and compare the results with the ORI method, we conduct a simulation study with varying sample sizes and number of time points. In each simulation, N subjects are generated, and each subject is observed at T time points and J pTyr sites. We assign the true profile for each of the J pTyr sites as follows. When $T = 4$, the 20% of sites are assigned to the null profile C_0, then 5% to profile C_1, and similarly, 10%, 10%, 5%, 15%, 5%, 15% and 15% of sites are assigned to profiles C_2 to C_8, respectively. The trajectories μ_{jt}s are randomly sampled from the sequence from -3 to 5 with increments 0.5 according to these pre-specified profiles. Data are generated based on model (2.1) with the error variance σ^2 set to 0.45. The variance components σ_η^2, σ_γ^2, and σ_δ^2 are all set as 0.1. The temporal correlation coefficient ρ is set as 0.5. 200 simulated datasets are generated and analyzed using both ORI algorithm and our Bayesian model.

2.4.1 Assessing the effect of sample size

The first simulation study is run with $N = 5$, $J = 35$, $T = 4$, and the time points are chosen as 0, 5, 30, and 60 minutes, which is similar to the setting of ovarian tumor data. Another simulation study is carried out with an increased sample size of $N = 20$. We draw 200 simulated datasets for each simulation setting. To evaluate the performance of our procedure, we consider the following four quantities: Type I error rate, Type II error rate, the overall correct classification proportion, and the accuracy of trajectory estimation. To measure the accuracy of the estimation for μ, we use a sum of squared error (SSE) loss function $L(\mu, \hat{\mu}) = \sum_{j=1}^{J} \sum_{t=1}^{T} (\hat{\mu}_{jt} - \mu_{jt})^2$, based on the posterior mean of $\hat{\mu}_{jt}$. All accuracy measurements are averaged over 200 simulated datasets.

Table 2.4 shows the simulation results of the two scenarios introduced above with $N = 5$ and $N = 20$. For the simulation with $N = 5$, the overall mean proportion of correct classification using the ORI method is 68.3%, compared to 94.1% from our Bayesian model using both the MAP and MPW estimators, corresponding to roughly 9 additional correctly assigned sites. The probabilities of making a Type I error $P(\mu_j \in \cup_{k=1}^{c-1} \hat{C}_k | \mu_j \in C_0)$ in our proposed model and ORI methods are 4% and 5%, respectively. The Type II error rates $P(\mu_j \in \hat{C}_0 | \mu_j \in \cup_{k=1}^{c-1} C_k)$ are around 0.2% for both the MAP and MPW estimators, compared to 0.4% based on the ORI method, indicating that all methods have high power in this scenario. The measurement for the accuracy of the estimated trajectory (lower SSE corresponds to a greater accuracy) is 18.5 for the Bayesian model compared to 54.1 for the ORI method, indicating that our model more accurately estimates the trajectory μ_j than the ORI method.

For the simulation with increased sample size with $N = 20$, the mean of the correct classification proportions of pTyr sites using the ORI method is 71.2% compared to 99.5% using both the estimators of our Bayesian model. Both of our proposed model and the ORI method accurately assign sites to the null, while our

Bayesian model does better in distinguishing the non-null profiles. In this scenario, our Bayesian model also performs better for estimating the trajectory than the ORI method.

We also construct the 75% credible sets in both the simulation settings. The proportions that the true profile is included in the 75% credible set are 98.5% and 99.9% for $N = 5$ and 20, respectively. Thus, we have confidence that in the rare case when the estimated profile is incorrect, the true classification is almost always in the credible set.

2.4.2 Assessing the effect of the signal-to-noise ratio and number of time points

We conduct additional simulation studies to examine the performance of our model under different signal-to-noise ratios and different numbers of time points. In the following simulation scenarios, we let $N = 20$ and $J = 35$ and consider three settings for T: $T = 4$, 5, and 8. For $T = 4$, each site is assigned to one of the 9 candidate profiles as in the previous section. However, as T increases, the number of candidate profiles increases exponentially. This is particularly challenging for the ORI method because its algorithm involves a recursive estimation scheme for each profile under consideration. To that end, we apply the ORI method to a restricted set of profiles, which are known to contain the set of true profiles. For $T = 5$, 11 out of 17 candidate profiles (including the null) are selected, representing 65% of the total number of possible profiles. For $T = 8$, the restricted set contains 15 out of 129 profiles, roughly 12% of all possible profiles. Table A1.3 in the Appendix reports the selected profiles and the percentage of the true assignment for the cases when $T = 5$ and 8.

In addition to examining the impact of T on accuracy, we also consider changes to the signal-to-noise ratio (i.e., effect size). For the largest effect size, μ_{jt} are drawn uniformly from the sequence from -3 to 5 with an increment of 0.1, subject to the profile membership. The medium effect size draws μ_{jt} from the sequence from -1.5

to 2.5 with an increment of 0.05, and the smallest effect size uses -0.75 to 1.25 with an increment of 0.025. The ORI method is applied to find the best profile among the selected set of the candidate profiles. To determine the accuracy without any additional knowledge of the restricted profile set, the MAP and MPW classifications are found as before. To compare against the ORI estimator that uses the selected profile set, we develop corresponding versions of the MAP and MPW estimators that use this extra knowledge. For the MAP estimator, we choose the candidate profile which has the maximum among $Pr(\mu_j \in C_r|Y)$ among r in the selected profile set. For the MPW estimator, we choose the profile from among the selected set with the maximum value of $\prod_{t=1}^{T-1} Pr(\mu_{jt} \square_t \mu_{j,t+1}|Y)$, where $\square_t$ is the t^{th} inequality in C_r.

Figure 2.2 shows the boxplots of the correct classification rates under the different simulation scenarios. In the $T = 4$ case, the mean correct classification rates of both the MAP and MPW methods are higher than the ORI method under all effect sizes. As the number of time points increases and/or the effect size decreases, correct classification becomes more challenging, and the accuracy deteriorates for all methods. However, across all scenarios the Bayesian methods that use the selected set of profiles consistently beat (often by a large margin) the ORI in accuracy. In fact, even without using the extra information on the restricted set of profiles, our Bayesian classification is often more accurate than the ORI (e.g., for all the effect sizes for $T = 4$ and 5, and for the large effect size for $T = 8$). Comparing the two estimators, we find that the MAP and MPW estimators are roughly equivalent for the small number of time points, and the MPW estimator is slightly better for the larger number of time points, which is consistent with the intuition discussed in Section 2.2.2. See Table 2.5 for detailed results of each simulation scenario. Figure 2.3 shows the boxplots of the accuracy of the estimated trajectories measured by SSE. In all cases, the Bayesian model performs better than the ORI method in estimating the trajectory μ_{jt}.

2.5 Discussion and Conclusion

In this project we develop a Bayesian methodology to examine the stability of tyrosine phosphorylated proteins undergoing cold ischemia and to characterize the direction of changes in the unstable ones. From the ovarian tumor data, 15 out of 32 phosphorylated proteins show significant changes during one-hour cold ischemia based on our Bayesian model, indicating that any scientific results related to these phosphorylated proteins may be altered by a short freezing delay after the tumor specimen excision.

A number of phosphorylated proteins that we identified as fluctuating in phosphorylation abundance under cold ischemia shock have been found to be medically important for therapy development or biomarker discovery in the literature. For example, DOK1 has been identified as a candidate tumor suppressor gene for various human malignancies (Lee et al., 2007; Berger et al., 2010) and works in a signal transduction pathway downstream of receptor tyrosine kinases (Némorin et al., 2001). This phosphorylated protein shows an increasing trend in phosphorylation abundances undergoing one hour cold ischemia shock in our study. GAB1, which serves in a different signaling pathway, was classified to the profile with an increase in phosphorylation abundance during the first 30 minute cold ischemia and a decrease during the next 30 minute cold ischemia based on our Bayesian model; it was also classified to the cluster characterized by a rapid increase in phosphorylation within 5 minutes of cold ischemia in Gajadhar et al. (2015). It has been found that GAB1 is usually over-expressed in cancer cells and may be used as a target for cancer therapy, especially for triple negative breast cancer patients (Chen et al., 2015; Gu and Neel, 2003). Similarly, EPHA2 has been found to be over-expressed in various cancers, particularly a high level of EPHA2 is detected in malignant cancer-derived cell lines and advanced forms of cancer (Tandon, Vemula and Mittal, 2011). EPHA2 displayed a decreasing trend in phosphorylation abundance during one hour cold ischemia based on our approach

and was also classified to the gradually decreased cluster in Gajadhar et al. (2015). It should be noted that PEAK1 may require a more careful sample collection protocol compared to the other phosphorylated proteins classified to the null profile. While the most probable explanation is that PEAK1 is not impacted by freezing delay, the non-null profile $C_6 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 > \mu_4\}$ is contained in the credible set, so the potential impact of cold ischemia should not be totally discounted.

Based on the simulation studies, our Bayesian model is more efficient for the classification of short time-course data than the ORI algorithm. Both our Bayesian model and the ORI method have high correct classification rates for the null profile C_0. However, our Bayesian model has a much higher accuracy in correctly classifying the non-null profiles than the ORI method. The Bayesian approach performs well even in the cases with more time points and small effect sizes. In addition to the improved accuracy, the computation speed of our Bayesian approach is much faster than the ORI method. In the simulation study with $T = 8$, it takes approximately 6 hours per dataset to use the ORI method on the subset of 15/129 profiles, whereas the Bayesian approach which considers all profiles requires only a third of the time.

Since the primary focus is whether the protein phosphorylation status has changed during one-hour cold ischemia, we only consider complete equality or a sequence of strict inequalities between each pair of time points. If the changes in phosphorylation abundance under a certain threshold need to be considered as equal, some extensions of our model can be considered. Wu et al. (2007) provides a similar approach using a 3-stage Markov model using the relationships $<$, $>$ and $=$ between two time points. However, they do not consider the role of the site or temporal dependence. As a simple ad hoc version, one might extend our method to allow the profiles with relationships $<$, $>$ and $=$ by choosing a cut-off value for approximate equality either through biological considerations or a default value. The pair μ_{jt} and $\mu_{j,t+1}$ would be considered to be approximately equal if $|\mu_{jt} - \mu_{j,t+1}| < \epsilon$, increasing

if $\mu_{jt} + \epsilon < \mu_{j,t+1}$, and decreasing if $\mu_{jt} - \epsilon > \mu_{j,t+1}$, where ϵ is a pre-specified positive value. A MAP or MPW estimator from the set of 3^{T-1} profiles can be found.

A key concern for an analysis with few patients is the validity of the normality assumption. A Q-Q plot and histogram of the standardized residuals ($Y_{ijt} - \hat{\mu}_{jt} - \hat{\gamma}_i - \hat{\delta}_{it} - \hat{\eta}_{ij}$) can be found in Figure 2.4. The Q-Q plot indicates that the normality assumption is reasonably satisfied in our analysis of the ovarian tumor data. If normality was found to be suspect, a transformation of the data (e.g., logarithm or Box-Cox) or a thicker-tailed or skewed distribution for ϵ_{ijt}, such as the t-distribution or skewed normal distribution, could easily be incorporated.

The phosphotyrosine signaling networks can be affected by cold ischemia shock for many phosphorylated proteins, even in a short time period. Our study shows that nearly half of the selected phosphorylated proteins experience a dramatic change during one-hour cold ischemia, which underscores the necessity of freezing the tissue sample immediately after excision. Because the phosphorylation changes of some pTyr sites classified to C_0 may be dramatically different among patients (see MAPK14 in Figure 2.1), these sites may require further attention. The current work provides an efficient method to detect unstable phosphorylated proteins under short-time cold ischemia shock and profile the direction of their changes. Thus, the results obtained from our method may provide valuable guidance on developing sample collection protocols and further analytical strategies.

2.6 Tables and Figures

Table 2.1: Detailed Result of Bayesian classification based on MAP. The probability of each protein being classified to the null profile is reported in 4^{th} column with the column title "$\hat{P}(Z_j = 0|Y)$". The posterior probabilities of classifying each protein to its corresponding profile is reported in the 5^{th} column with the column title "Maximum Posterior Probabilities". The candidate profiles included in 75% credible set for each protein are reported in the last column. * indicates that the pTyr site is classified to the same profile using both the ORI method and the Bayesian model.

Protein	pTyr	Sequence	$\hat{P}(Z_j = 0\|Y)$	Maximum Posterior Probabilities	75% Credible Set
$C_0 = \{\mu \in R^4 : \mu_1 = \mu_2 = \mu_3 = \mu_4\}$					
annexin A2 isoform 2	ANXA2*	LSLEGDHSTPPSA[y]GSVK	0.970	0.970	C_0
glucocorticoid receptor DNA binding factor 1	ARHGAP35*	NEEENI[y]SVPHDSTQGK	0.979	0.979	C_0
cell division cycle 2 protein isoform 1	CDK1*	IGEGT[y]GVVYK	0.951	0.951	C_0
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A isoform 3	DYRK1A	IYQ[y]IQSR	0.984	0.984	C_0
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2 isoform 1	DYRK2*	VYT[y]IQSR	0.974	0.974	C_0
ephrin receptor EphB3 precursor	EPHB3*	FLEDDPSDPT[y]TSSLGGK	0.850	0.850	C_0
filamin A, alpha isoform 1	FLNA*	VHSPSGALEEcYVTEIDQDK[y]AVR	0.875	0.875	C_0
high density lipoprotein binding protein	HDLBP*	MD[y]VEINIDHK	0.867	0.867	C_0
homeodomain-interacting protein kinase 1 isoform 2	HIPK1*	AVcST[y]LQSR	0.980	0.980	C_0
NKF3 kinase family member	PEAK1	ASTDVAGQAVTINLVPTEEQAKP[y]R	0.523	0.523	C_0,C_6
phosphoinositide-3-kinase, regulatory subunit 2 (beta)	PIK3R2*	SREYDQL[y]EEYTR	0.968	0.968	C_0
serine/threonine-protein kinase PRP4K	PRPF4B*	LcDFGSASHVADNDITP[y]LVSR	0.986	0.986	C_0
splicing factor, arginine/serine-rich 9	SRSF9*	G[s]PHYFSPFRPY	0.974	0.974	C_0
signal transducer and activator of transcription 3 isoform 1	STAT3	YcRPESQEHPEADPGSAAP[y]LK	0.903	0.903	C_0
tensin like C1 domain containing phosphatase isoform 1	TENC1*	GPLDGSP[y]AQVQRPPR	0.970	0.970	C_0
tensin	TNS1*	DDGMEEVVGHTQGPLDGSL[y]AK	0.958	0.958	C_0
vinculin isoform VCL	VCL*	SFLDSG[y]R	0.956	0.956	C_0
$C_1 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 < \mu_4\}$					
docking protein 1	DOK1	VKEEGYELPYNPATDD[y]AVPPPR	0.069	0.506	C_1,C_2,C_5
SH2 containing inositol phosphatase isoform b	INPP5D	EKL[y]DFVK	0.000	0.721	C_1,C_2
myelin protein zero-like 1 isoform a	MPZL1*	SESVV[y]ADIR	0.077	0.586	C_1,C_5
$C_2 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 > \mu_4\}$					
GRB2-associated binding protein 1 isoform a	GAB1*	SSGSGSSVADERVD[y]VVVDQQK	0.000	0.996	C_2
Wiskott-Aldrich syndrome gene-like protein	WASL*	VI[y]DFIEK	0.001	0.851	C_2
$C_3 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 < \mu_4\}$					
mitogen-activated protein kinase 14 isoform 1	MAPK14*	HTDDEMTG[y]VATR	0.000	0.771	C_3
$C_4 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 > \mu_4\}$					
paxillin	PXN	VGEEEHV[y]SFPNK	0.000	0.666	C_4,C_3
$C_5 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 < \mu_4\}$					
mitogen-activated protein kinase 7 isoform 1	MAPK7	GLcTSPAEHQYFMTE[y]VATR	0.000	0.402	C_5,C_1
$C_6 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 > \mu_4\}$					
TYRO protein tyrosine kinase binding protein isoform 1 precursor	TYROBP	ITETESP[y]QELQGQR	0.044	0.946	C_6
$C_7 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 < \mu_4\}$					
NCK adaptor protein 1	NCK1*	L[y]DLNMPAYVK	0.000	0.820	C_7
neural precursor cell expressed, developmentally down-regulated 9 isoform 1	NEDD9	EKD[y]DFPPPMR	0.010	0.463	C_7,C_3,C_8
$C_8 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 > \mu_4\}$					
cyclin-dependent kinase-like 5	CDKL5*	NLSEGNNANYTE[y]VATR	0.020	0.668	C_8,C_6
ephrin receptor EphA2	EPHA2*	VLEDDPEAT[y]TTSGGKIPIR	0.000	0.844	C_8
FYN binding protein (FYB-120/130) isoform 2	FYB*	TTAVEID[y]DSLK	0.474	0.254	C_0,C_8,C_4
protein kinase C, delta	PRKCD*	RSDSASSEPVGI[y]QGFEK	0.000	0.419	C_8,C_4

Table 2.2: Posterior median and 95% CI for each model parameter

	Posterior Median	95% CI
σ	0.173	(0.164, 0.185)
μ_0	1.002	(0.929, 1.078)
σ_μ	0.278	(0.228, 0.347)
$\sigma_{\mu*}$	0.104	(0.080, 0.141)
σ_γ	0.017	(0.002, 0.076)
σ_η	0.003	(0.000, 0.017)
σ_δ	0.107	(0.076, 0.161)
ρ	0.131	(0.005, 0.504)
θ	0.487	(0.309, 0.673)

Table 2.3: Comparison of classifications from the ORI and MAP estimators for the human tumor data.

		ORI Method								
		C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Bayesian Approach (MAP)	C_0	14	0	1	0	0	0	2	0	0
	C_1	0	1	2	0	0	0	0	0	0
	C_2	0	0	2	0	0	0	0	0	0
	C_3	0	0	0	1	0	0	0	0	0
	C_4	0	0	0	0	0	0	0	0	1
	C_5	0	0	0	0	0	0	1	0	0
	C_6	1	0	0	0	0	0	0	0	0
	C_7	0	0	0	0	0	0	0	1	1
	C_8	0	0	0	0	0	0	0	0	4

Table 2.4: Simulation results from Section 2.4.1. 200 simulated datasets with $J = 35$ and $T = 4$ were used in both the simulation settings. In each simulation, 8 pTyr sites were assigned to the null profile C_0 and 27 pTyr sites were assigned to the other candidate profiles. The quantities reported in this table are averaged across 200 simulated datasets. The accuracy of the estimated trajectory is measured by the sum of squared errors (SSE), and a lower value indicates a higher accuracy.

	$N = 5$			$N = 20$		
	Bayesian		ORI	Bayesian		ORI
	MAP	MPW		MAP	MPW	
Correct classification (overall)	94.1%	94.1%	68.3%	99.5%	99.5%	71.2%
Type I error rate: $Pr(\mu_j \in \cup_{k=1}^{c-1} \hat{C}_k \mid \mu_j \in C_0)$	4.0%	4.0%	5.0%	0.6%	0.6%	0.0%
Type II error rate: $Pr(\mu_j \in \hat{C}_0 \mid \mu_j \in \cup_{k=1}^{c-1} C_k)$	0.2%	0.2%	0.4%	0.0%	0.0%	0.4%
Trajectory Accuracy: $\sum_{j=1}^{J} \sum_{t=1}^{T} (\hat{\mu_{jt}} - \mu_{jt})^2$	18.5	18.5	54.1	4.3	4.3	38.5

Table 2.5: Simulation results from section 2.4.2. The correct classification rates is formed under different T and different effect sizes. For each simulation study, $N = 20$, $J = 35$ and 200 simulated datasets were used. The accuracy of the estimated trajectory is measured by the sum of squared errors.

		$T = 4$			$T = 5$					$T = 8$				
		All Possible CPs			All Possible CPs		Selected CPs			All Possible CPs		Selected CPs		
		MAP	MPW	ORI	MAP	MPW	MAP	MPW	ORI	MAP	MPW	MAP	MPW	ORI
Correct classification (overall)	Large Effect Size	94.8%	94.9%	68.6%	91.4%	91.6%	95.0%	95.1%	77.4%	83.6%	84.3%	97.1%	97.3%	72.5%
	Medium Effect Size	87.3%	87.5%	63.4%	80.9%	81.5%	89.0%	89.3%	73.1%	66.6%	68.1%	92.9%	93.5%	70.1%
	Small Effect Size	71.5%	71.8%	41.6%	61.2%	63.3%	75.4%	76.0%	56.2%	45.8%	49.1%	82.6%	85.7%	64.6%
Type I Errors	Large Effect Size	0.6%	0.6%	0.0%	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
	Medium Effect Size	4.2%	4.2%	0.0%	1.3%	1.3%	1.3%	1.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
	Small Effect Size	28.3%	28.3%	0.0%	14.3%	14.3%	14.3%	14.3%	0.0%	1.4%	1.4%	1.4%	1.4%	0.1%
Type II Error	Large Effect Size	0.0%	0.0%	1.9%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%
	Medium Effect Size	1.0%	1.0%	12.9%	0.3%	0.3%	0.3%	0.3%	2.2%	0.0%	0.0%	0.0%	0.0%	0.2%
	Small Effect Size	3.4%	3.4%	59.7%	1.8%	1.8%	1.8%	1.8%	29.2%	0.1%	0.1%	0.1%	0.1%	3.6%
Trajectory Accuracy	Large Effect Size	4.2		33.8	5.6		21.2			8.6				147.5
	Medium Effect Size	4.2		13.7	5.6		9.9			8.5				42.6
	Small Effect Size	4.3		15.8	5.9		11.6			8.3				17.5

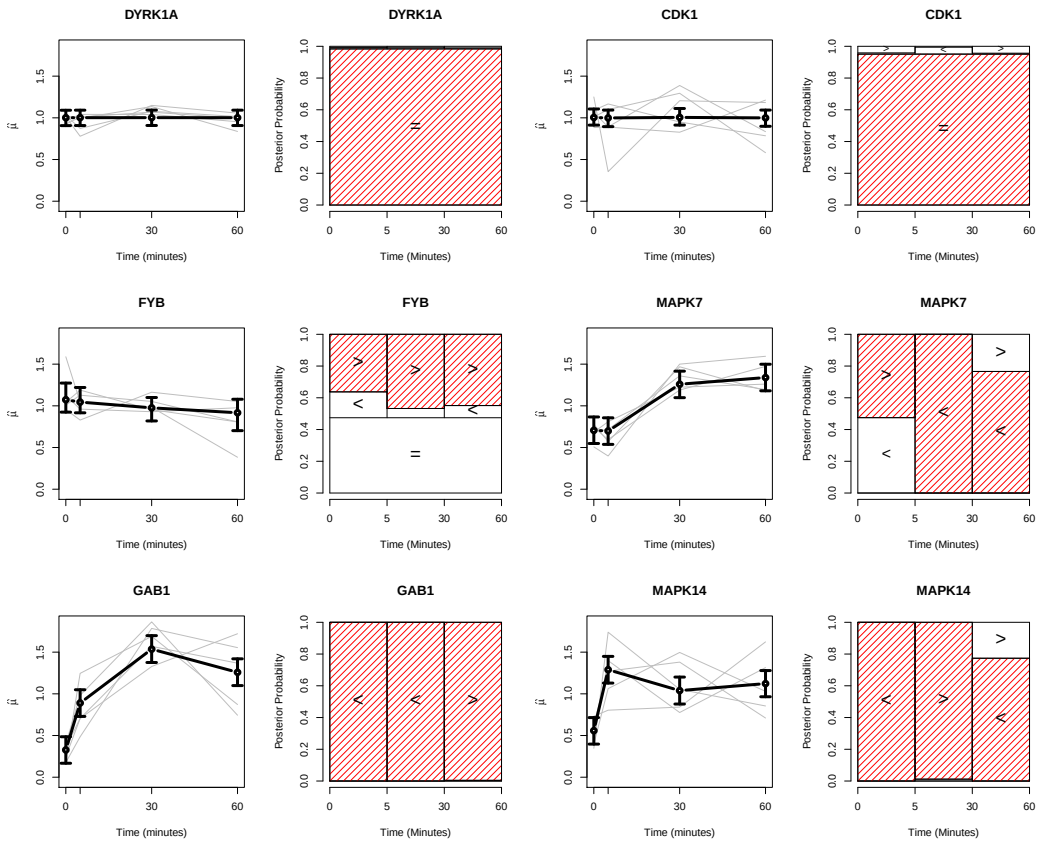

Figure 2.1: Estimated trajectory with 95% credible interval (1^{st} and 3^{rd} columns) and MPW estimators (2^{nd} and 4^{th} columns) for six representative phosphorylated proteins. The gray lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

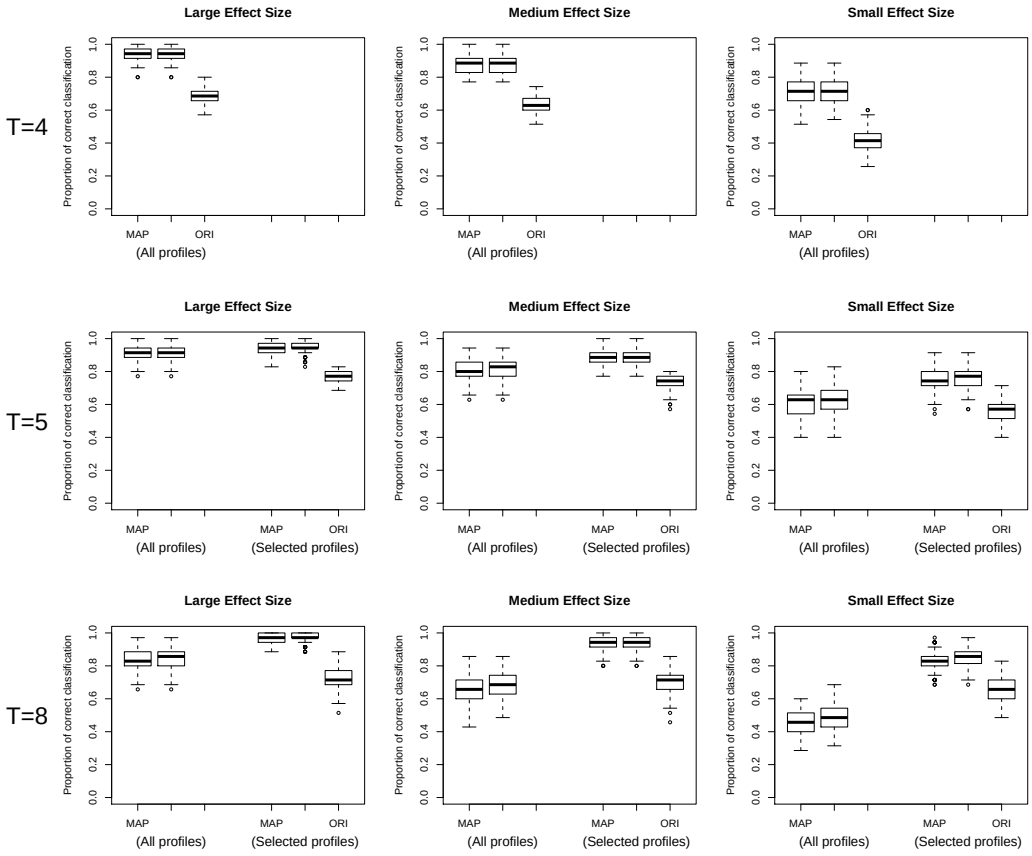

Figure 2.2: Boxplots of simulation results from Section 2.4.2. The correct classification proportions shown are calculated based on 200 simulated datasets for each case. For $T = 4$, all the possible candidate profiles are considered. For $T = 5$ and 8, the ORI method is not used when considering all possible profiles due to the limitation of computation speed.

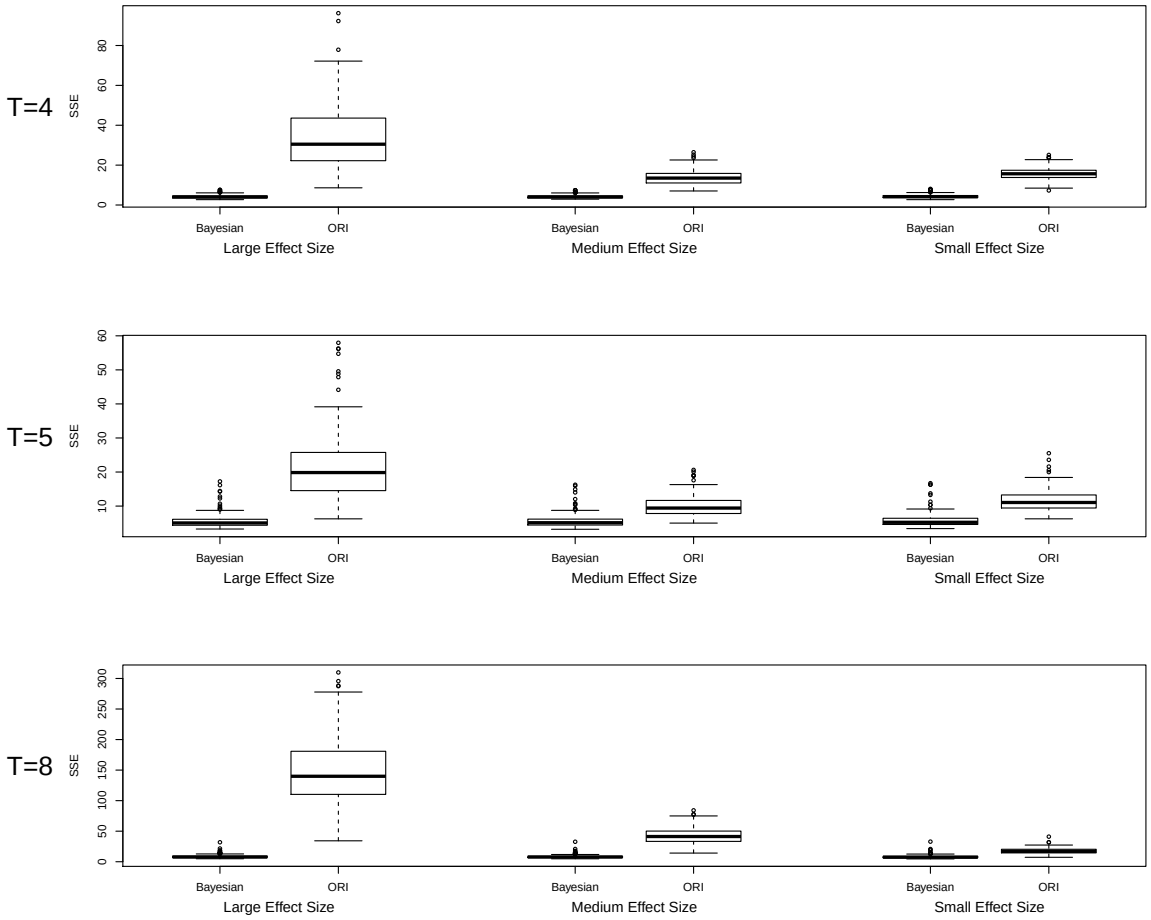

Figure 2.3: Boxplots of sum of squared error (SSE) measuring the accuracy of trajectory estimation. A lower value indicates higher accuracy.

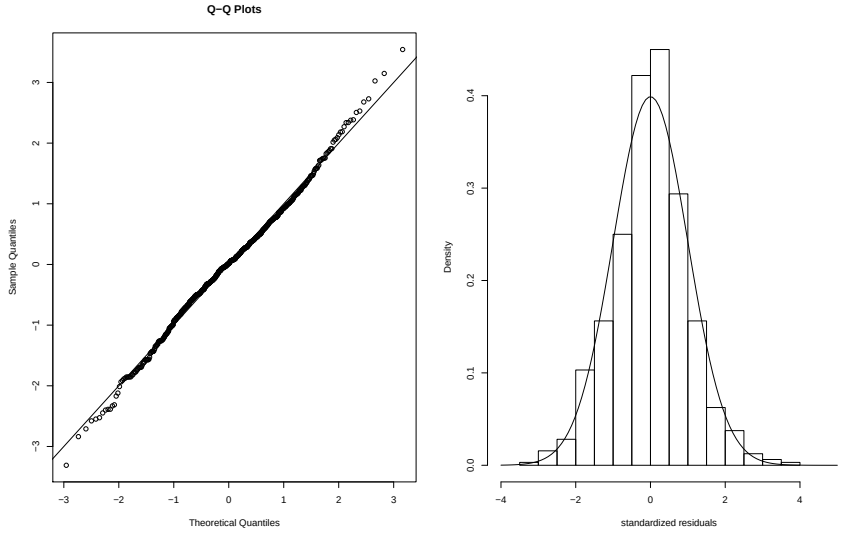

Figure 2.4: Q-Q plot and histogram of estimated residuals for ovarian tumor data.

CHAPTER 3

INVESTIGATION OF DIFFERENT STATISTICAL METHODS FOR ESTIMATING TREATMENT EFFECTS WHEN OUTCOME VARIABLE IS ORDINAL AND CONFOUNDING EXISTS

3.1 Introduction

Ordinal categorical data has been quite commonly used in the clinical, social and economic sciences (Agresti, 1996). For example, the commonly used Likert scale in social survey study is an ordinal categorical variable with five different categories: "Strongly disagree", "Disagree", "Neither agree nor disagree", "Agree" and "Strongly agree", to capture the attitude of a subject to a certain statement (Likert, 1932). The ordinal measured variables are often used as the primary outcome variable in clinical trials. For example, the World Health Organization (WHO) uses an ordinal outcome with four categories as the primary outcome to examine the efficacy of antimalarial drugs for uncomplicated malaria: 1) ETF (early treatment failure); 2) LCF (late clinical failure); 3) LPF (late parasitological failure) and; 4) ACPR (adequate clinical and parasitological response) (Whegang et al., 2010). In our case study, we examine whether physical activity contributes to diabetes status, with adjustment for subject's demographic information and dietary information. The outcome variable is classified as diabetes, pre-diabetes, metabolic syndrome and normal. Studies with ordinal outcome variable are common, and it is important to use the analytical methods which control the unbalanced covariates and have good statistical properties.

When the data arise from a randomized control trial, the subjects are randomly assigned to different groups, say treatment and comparator groups. It is generally assumed that there are no confounding covariates, either measured or unmeasured (Austin, 2011). The commonly used methods for analyzing ordinal outcome variable include parametric statistic method such as the cumulative logit/probit methods (Agresti, 1996) and non-parametric method, such as the Mann-Whitney U test statistics (Mann and Whitney, 1947). The cumulative logit/probit models link the cumulative probability and linear combination covariates with different intercepts for different levels of responses (Agresti and Kateri, 2016). The parametric approach provides a powerful tool if the model is correctly specified. The non-parametric method such as the Mann-Whitney U test statistic is also a powerful way to examine the treatment effects for the ordinal outcomes when there is no confounding variables (Mann and Whitney, 1947; Grissom, 1994). However, when the data arise from natural settings such as electronic clinical records or registry data, the treatment selection may depend on the patients' own characteristics. The covariates between treatment and comparator group may not be balanced. The statistical methods without considering the selection bias may result in biased estimates for treatment effect. Historically, regression adjustment methods such as cumulative logit/probit model is often used to account for the difference in baseline characteristics between treatment and comparator groups. Recently, the propensity score based methods, such as inverse probability weighting, stratification and matching, have been widely used to reduce the impact from the unbalanced confounding covariates in observational data (Austin, 2011; Abdia et al., 2017). These propensity score based methods are free of outcome-specific model and thus are robust in estimation of the treatment effect. These methods may be applied to compare treatment effect for ordinal outcome variable by assigning numeric scores to the ordinal outcomes, thus loosing the ordinal nature of the ordinal variables.

In the literature, the superiority score has been considered as a measure for treatment effect for ordinal outcome (Klotz, 1966; Ryu and Agresti, 2008; Agresti and Kateri, 2016). The Mann-Whiteney U test statistic can be considered as a superiority score although the confounding variables are not adjusted (Mann and Whiteny, 1947). In order to address the confounding covariates in the observational data, Glen, Kong and Datta (2017) developed an adjusted U-statistic to test the equality of distributions across multiple groups when the confounding covariates exist. Agresti and Kateri (2016) presented probit/logit model based methods to estimate the superiority score. However, the parametric models such as probit/logit models are prone to provide biased estimates if the model is misspecified. On the other hand, the propensity score based methods such as stratification, matching, and weighting may provide excellent way to estimate treatment effect with control of confounding variables. The propensity score is also called balance score and the propensity score based methods usually do not assume outcome regression models. In this project, we investigate the confounding-adjusted to estimate the superiority score for ordinal outcomes with control of confounding covariates.

This chapter is organized as following: In Section 3.2, we investigate the statistical methods for assessing treatment effects for ordinal outcome when confounding variables exist. These methods include the cumulative logit model, propensity score based matching and stratification, and the adjusted U-statistic. In Section 3.3, simulation studies are carried out to compare the performance of these methods. In Section 3.4, a case study using the National Health and Nutrition Examination Survey (NHANES) data is carried out to examine the causal effect of physical activities on diabetes status with adjustment of demographic and dietary information. The last section is devoted to a discussion.

3.2 Statistical Method

Suppose N subjects are included in the sample, and X_i, Y_i, T_i, respectively, denotes a vector of p confounding variables, the ordinal outcome variable, and the treatment assignment for i^{th} subject in the study. Here Y_i is an outcome variable with C ordinal categories (i.e., $Y_i \in \{1, 2, \ldots, C\}$), and T_i takes 1 if the i^{th} subject is assigned to treatment group, and takes 0 if the i^{th} subject is assigned to comparator group. For notational convenience, numeric numbers are assigned to the outcome variable Y_i to keep the ordering but without numeric meanings. For the i^{th} subject, there are two potential outcomes: the potential outcome under treatment $Y_i^{(1)}$ and the potential outcome under comparator group $Y_i^{(0)}$. The observed outcome is:

$$Y_i = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)} = \begin{cases} Y_i^{(0)}, \text{ if } T_i = 0, \\ Y_i^{(1)}, \text{ if } T_i = 1. \end{cases}$$

Here $Y_i^{(0)}$ and $Y_i^{(1)}$ are ordinal categorical variables with C categories, a higher category indicates a better outcome. Here we assume that all confounding variables are observed, and the assumptions on overlap and strong ignorability (Rosenbanm and Robin, 1983) hold. For ordinal outcome variable, the stochastic superiority score (Klotz, 1966) measures if the outcome under treatment is stochastically larger than the outcome under comparator group. The superiority score is defined as $Pr(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2} Pr(Y_i^{(1)} = Y_i^{(0)})$. Under the null hypothesis that there is no treatment effect, the superiority score equals to 0.5. In the following subsection 3.2.1, we investigate the parametric method to estimate the superiority score. On Section 3.2.2 we apply the propensity score-based matching and stratification to the superiority score, and in Section 3.2.3 we apply the adjusted Mann Whitney U test statistics to assess the superiority score when the confounding covariates exist.

3.2.1 Parametric method for estimating superiority score

The ordinal logistic regression (OLR) is also known as the cumulative logit model. It is a widely used method to estimate the treatment effect for ordinal outcome variable. The OLR assumes that the effect of the treatment (or covariates) is identical for all $C-1$ cumulative logits and a single parameter is used to describe the treatment effect when the model fits data well (Agresti, 1996). The OLR model uses a set of dichotomized logistic models as follows:

$$log\frac{P(Y \leq k|X,T)}{P(Y > k|X,T)} = \alpha_k - X\delta - \tau T,$$

where $k = 2, \ldots, C$. The parameters $(\alpha_1, \ldots, \alpha_k, \delta, \tau)$ can be estimated based on the maximum likelihood estimation. Hypothesis testing for treatment effect is conducted by the Wald test for testing the hypothesis that $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$. The cumulative logit model uses the proportional odds structure to interpret the treatment effects. However, the superiority score can be approximated from the maximum likelihood estimate of τ (Agresti and Kateri, 2016):

$$\hat{\gamma}_{OLR} \approx \frac{exp(\hat{\tau}/\sqrt{2}))}{1+exp(\hat{\tau}/\sqrt{2})}.$$

The estimated variance of $\hat{\gamma}_{OLR}$ is calculated based on delta-method: $\hat{\sigma}^2_{\gamma,OLR} = \frac{exp(\sqrt{2}\hat{\tau})}{2[1+exp(\hat{\tau}/\sqrt{2})]^4} \cdot \hat{\sigma}^2_{\tau}$. The hypothesis testing for treatment effect is equivalent to test $H_0 : \gamma_{OLR} = 0.5$ vs. $H_1 : \gamma_{OLR} \neq 0.5$. The test can be carried out using the Wald statistic $z = \frac{\hat{\gamma}_{OLR}-0.5}{\hat{\sigma}_{OLR}}$, where $\hat{\sigma}^2_{OLR}$ is the estimated variance for $\hat{\gamma}_{OLR}$, and the p-value is obtained from $2\Phi(-|z|)$, where Φ is the cumulative density function of a standard normal distribution.

3.2.2 Propensity score based methods for estimating superiority score

The propensity score is the probability of receiving treatment conditional on observed baseline covariates (Rosenbaum and Rubin, 1983), which can be written as:

$$e_i(X_i) = Pr(T_i = 1|X_i),$$

where $i = 1, \ldots, N$. Rosenbaum and Rubin (1983) proved that similar distributions of propensity scores between treatment and comparator groups implies similar distributions of confounding covariates between treatment and comparator groups, thus propensity score is also considered as a balancing score. The propensity score could be estimated based on parametric model such as logistic regression model or machine learning method such as generalized boosting method (Abdia et al., 2017). However, in this project, the propensity score is estimated from the logistic regression model:

$$log\frac{e_i(X_i)}{1 - e_i(X_i)} = log\left\{\frac{Pr[T = 1|X_i]}{1 - Pr[T = 1|X_i]}\right\} = X_i\beta \tag{3.1}$$

The estimate of β is obtained by using the maximum likelihood method. Two propensity score based methods are considered: propensity score-based matching (PSM) and propensity score-based stratification.

1. Propensity score-based matching

Propensity score-based matching could be implemented by first obtaining the counterfactual outcomes via matched samples of the treated and comparative subjects, and then estimating the treatment effect by averaging the subject-level treatment effects over the N subjects (Rosenbaum & Rubin, 1983, 1985). The steps to implement propensity score-based matching are as follows:

Step 1: Calculate propensity score for each subject $e_i = Pr(T_i = 1|X_i)$, where $i =$

$1, \ldots, N$;

Step 2: For i^{th} subject, if $T_i = 1$, then $Y_i^{(1)}$ is observed, and $Y_i^{(0)}$ is estimated from a subject j in the comparator group such that the propensity score e_j is the closest one to e_i. Similarly, for a subject in comparator group (i.e., $T_i = 0$), $Y_i^{(0)}$ is observed, and the counterfactual outcome $Y_i^{(1)}$ is estimated from a subject in the treatment group whose propensity score is the closest one to e_i;

Step 3: The outcome $(Y_i^{(0)}, Y_i^{(1)})$ for i^{th} subject is thus able to be used to assess the subject-level treatment effect for i^{th} subject, that is, $\hat{\gamma}_i = I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2} I(Y_i^{(1)} = Y_i^{(0)})$;

Step 4: The stochastic superiority score is thus calculated based on the entire matched sample:

$$\hat{\gamma}_{Mat} = \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_i = \frac{1}{N} \sum_{i=1}^{N} [I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2} I(Y_i^{(1)} = Y_i^{(0)})].$$

The variance of $\hat{\gamma}_{Mat}$ can be estimated by:

$$\hat{\sigma}^2_{\gamma,Mat} = \tfrac{1}{N}(\hat{p}_1 + \tfrac{1}{4}\hat{p}_0 - \hat{\gamma}^2_{Mat})$$

where $\hat{p}_1 = \frac{1}{N} \sum_{i=1}^{N} I(Y_i^{(1)} > Y_i^{(0)})$ and $\hat{p}_0 = \frac{1}{N} \sum_{i=1}^{N} I(Y_i^{(1)} = Y_i^{(0)})$ (see Appendix 2 for the detailed derivation of $\hat{\sigma}_{\gamma,Mat}$). The hypothesis test $H_0 : \gamma_{Mat} = 0.5$ vs. $H_1 : \gamma_{Mat} \neq 0.5$ is carried out by using the Wald test statistic $z = \frac{\hat{\gamma}_{Mat} - 0.5}{\hat{\sigma}_{\gamma,Mat}}$ with p-value as $2\Phi(-|z|)$, where Φ is the cumulative density function of a standard normal distribution.

2. Propensity score based stratification

The propensity score-based stratification methods is to divide subjects to several different strata based on their propensity scores. Within each stratum, the distributions

of the covariates are considered similar for treated subjects and comparative subjects. Thus, the treatment effect can be estimated within each stratum, and the treatment effect over entire sample can be estimated by the average treatment effects across different strata. The steps to carry out the propensity score-based stratification are as follows:

Step 1: Estimate the propensity score $\hat{e}_i = \hat{P}(T_i = 1|X_i)$ for $i = 1, \ldots, N$;

Step 2: Rank the subjects with respect to their propensity scores, and then stratify all subjects to S strata (usually $S = 5$) based on the quantile of the propensity scores;

Step 3: Examine the covariate balance based on the absolute standardized mean difference (ASMD) (McCaffrey et al., 2013). For p^{th} covariate, the ASMD is calculated by:

$$ASMD_p = \frac{\sum_{s=1}^{S} \frac{n_s}{N} |\bar{X}_{p,s}^{(1)} - \bar{X}_{p,s}^{(0)}|}{sd_p}$$

where $\bar{X}_{p,s}^{(1)}$ and $\bar{X}_{p,s}^{(0)}$ are the mean of p^{th} covariate in the treatment group and comparative group, respectively. sd_p is the standard deviation of p^{th} covariate in the entire sample, and n_s is the sample size at the s^{th} stratum. The p^{th} covariate is considered to be balanced if $ASMD_p \leq 0.2$;

Step 4: Let X^{unbal} denote the unbalanced covariates identified from step 3. For s^{th} stratum, the treatment effect is estimated from the ordinal logistic regression:

$$log \frac{P(Y \leq k|X^{unbal}, T)}{P(Y > k|X^{unbal}, T)} = \alpha_k - X^{unbal}\delta - \tau_s T,$$

the parameters $(\alpha_k, \delta_i, \tau_s)$ are estimated by the maximum likelihood method based on the stratum-specific data $(Y_s, T_s, X_s^{unbal})(s = 1, \ldots, 5)$. According to Agresti and Kateri (2016), the superiority score for s^{th} stratum is estimated by $\hat{\gamma}_s \approx \frac{exp(\hat{\tau}_s/\sqrt{2}))}{1+exp(\hat{\tau}_s/\sqrt{2})}$, and its variance is estimated by $\hat{\sigma}^2_{\gamma,s} = \frac{exp(-\sqrt{2}\hat{\tau}_s)}{2[1+exp(-\hat{\tau}_s/\sqrt{2})]^4} \cdot \hat{\sigma}^2_{\tau_s}$;

Step 5: The overall superiority score is thus estimated by pooling all the stratum-specific estimates across all strata, which is $\hat{\gamma}_{Strat} = \sum_{s=1}^{S} \frac{n_s}{N} \hat{\gamma}_s$. The variance of the overall superiority score is then estimated by $\hat{\sigma}^2_{\gamma,Strat} = \sum_{s=1}^{S} (\frac{n_s}{N})^2 \hat{\sigma}^2_{\gamma,s}$;

Step 6: Hypothesis testing of $H_0 : \hat{\gamma}_{Strat} = 0.5$ vs. $H_0 : \hat{\gamma}_{Strat} \neq 0.5$ is carried out by using the Wald test statistic $z = \frac{\hat{\gamma}_{Strat} - 0.5}{\hat{\sigma}_{\gamma,Strat}}$ with p-value $2\Phi(-|z|)$ as stated before.

3.2.3 Adjusted Mann-Whitney U test statistic

The Mann-Whitney U statistic (Mann and Whitney, 1947) is powerful to examine treatment effect between two groups when there is no confounding covariates. The classic Mann-Whitney U statistics has the following form:

$$\gamma_U = \frac{1}{n_1 n_0} \sum_{\boldsymbol{i} \in \{i:T_i=1\}} \sum_{\boldsymbol{j} \in \{j:T_j=0\}} K(Y_i, Y_j) \tag{3.2}$$

where n_1 and n_0 are the numbers of subjects in treatment and comparator groups, respectively. When the Wilcoxon kernel $K(Y_i, Y_j) = I[Y_i > Y_j] + \frac{1}{2} I[Y_i = Y_j]$ is used, this U statistic naturally reflects the stochastic superiority score. If the distributions of outcome under treatment and comparator conditions are identical and there is no confounding covariates, the superiority score equals 0.5. However, when there are confounding covariates, two sample U statistics may not be appropriate to assess the treatment any more.

In the presence of confounding covariates X, Glen, Kong and Datta (2017) proposed an adjusted form of the U-statistic using the inverse probability weighting to weight each subject into its study population. The weights are calculated based

on the propensity scores. That is, the weight for i^{th} observation is calculated as

$$w(x_i, t_i\,; \boldsymbol{\beta}) = \begin{cases} \frac{1}{e_i(X_i)}, \text{ if } T_i = 1, \\ \frac{1}{1-e_i(X_i)}, \text{ if } T_i = 0. \end{cases} \tag{3.3}$$

Here, $e_i(X_i)$ is the propensity score estimated from Equation 3.1. The adjusted U-statistic is defined as the following form:

$$\gamma_U = \frac{1}{n_1}\frac{1}{n_0} \sum_{\boldsymbol{i}\in\{i:T_i=1\}} \sum_{\boldsymbol{j}\in\{j:T_j=0\}} \left\{ \widetilde{w}_1(X_i, T_i; \widehat{\boldsymbol{\beta}}) \right\} \times K(Y_i, Y_j) \times \left\{ \widetilde{w}_0(X_j, T_j; \widehat{\boldsymbol{\beta}}) \right\}$$

Here the Wilcox kernel $K(Y_i, Y_j)$ is applied, and the weight for i^{th} subject $\widetilde{w}_t(X_i, T_i; \widehat{\boldsymbol{\beta}})$ is a normalized weight which is obtained from

$$\widetilde{w}_t(X_i, T_i; \widehat{\boldsymbol{\beta}}) = \frac{w(X_i, T_i; \widehat{\boldsymbol{\beta}})}{\widehat{W}_t(\widehat{\boldsymbol{\beta}})}, \text{ for } Ti = t.$$

Here, $w(X_i, T_i; \widehat{\boldsymbol{\beta}})$ is calculated based on the propensity score in Equation 3.3 and $\widehat{W}_t(\widehat{\boldsymbol{\beta}})$ is the averaged weight for group t:

$$\widehat{W}_t(\widehat{\boldsymbol{\beta}}) = \frac{1}{n_t} \sum_{i=1}^{N} w(X_i, T_i; \widehat{\boldsymbol{\beta}}) \cdot I(T_i = t)$$

where $t = 0, 1$.

In Glen, Kong and Datta (2017), the asymptotic normality of the adjusted U-statistics is given and a close form of the asymptotic variance is presented. The hypothesis for testing $H_0 : \gamma_u = 0.5$ vs. $H_1 : \gamma_u \neq 0.5$ is thus able to be carried out. Under the null hypothesis,

$$T = \frac{\gamma_u - 0.5}{\sqrt{\hat{v}}} \sim N(0, 1)$$

where $\hat{v}$ is the asymptotic variance of γ_u and is estimated from the observed data. The detailed formula of $\hat{v}$ can be found in Glen, Kong and Datta (2017).

3.3 Simulation Study

Simulation studies are carried out to compare the performance of different methods presented in Section 3.2. An ordinal outcome with four levels (say, 1, 2, 3, 4) is considered. It should be noticed that the numeric numbers assigned to the outcome are only for notational convenience, and these numeric numbers do not bear numerical meanings. The treatment assignment variable T takes values 0 and 1, where $T = 1$ indicates treated subjects and $T = 0$ indicates comparative subjects. Three different covariates $X = (X_{.1}, X_{.2}, X_{.3})^T$ are considered, where $X_{.1} \sim N(0, 1)$, $X_{.2} = X^*_{.2} - 0.5$, where $X^*_{.2} \sim Bin(1, 0.5)$ and $X_{.3} \sim Uniform(-0.5, 0.5)$. The treatment assignment T is generated based on the logistic regression model:

$$log\frac{P(T=1|X)}{P(T=0|X)} = X\beta.$$

The covariate parameter β is set as as $\beta = (1, 1, 1)$. Two scenarios are considered in the simulation study. Scenario 1 assumes that the underlying outcome variable follows an ordinal logistic regression, and scenario 2 assumes that the outcome variable follows a cumulative Box-Cox model (Guerrero and Johnson, 1982). Since the ordinal logistic regression is powerful when the underlying model is correctly specified and is sensitive when the model is misspecified, we expect the ordinal logistic regression model have the best performance among all the methods under scenario 1 and this advantages disappears when the underlying model is misspecified.

3.3.1 Scenario 1: Outcome follows an ordinal logistic model

In this simulation scenario the outcome variable was generated based on the ordinal logistic regression model:

$$log\frac{P(Y \leq k|X,T)}{P(Y > k|X,T)} = \alpha_k + X\delta + \tau T$$

where $k = 1, 2, 3$. We set the intercepts $(\alpha_1, \alpha_2, \alpha_3) = (-1.39, -0.85, -0.62)$ so that the probability of occurrence for the first level of the outcome as 0.2, i.e., $Pr(Y = 1) = 0.2$, and the probabilities for the second to fourth levels are roughly set as 0.1, 0.05, and 0.65, respectively, to match the percentages of outcomes in the case study presented in next section. The covariate effect parameter δ is assigned as $\delta = (1, 1, 1)$. The treatment effect parameter τ is considered to have values from sequence 0 to 0.6 with 0.05 increment.

We generated 1000 simulated data for each τ, and each simulated data includes 5000 observations. All the methods presented in Section 3.2 were applied to estimate the treatment effect. The average of the rejection of the hypothesis test at a significance level 0.05 for the 1000 simulated data is calculated. For comparison convenience, the estimated treatment effect parameter in the ordinal logistic regression model is transformed to the superiority score using the approximation from Agresti and Kateri (2016). Under the null hypothesis that there is no treatment effect, the sizes of all methods are reported in Table 3.1. Only the ordinal logistic regression and the adjusted U statistic have the size close to 0.05. The propensity score-based matching, stratification and the classical Mann-Whitney test all have sizes much larger than 0.05, thus not suitable for estimating treatment effect. For power calculation, we only consider the methods with a size close to 0.05 as the valid approach. Thus, only the ordinal logistic regression and the adjusted U statistic are included in the power analysis.

Figure 3.1 shows the power curves of the ordinal logistic regression and the adjusted U statistics based on the simulated datasets from the ordinal logistic regression model. As expected, the ordinal logistic regression shows higher power than the adjusted Mann-Whitney U test statistics. Table 3.2 summarizes the simulation results in this scenario. The true superiority score γ is calculated as the average of the 1000 estimated superiority scores based on the 1000 simulated data. The superiority score for each data set was calculated from $\gamma = N^{-1}\sum_{i=1}^{N}[I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2}I(Y_i^{(1)} = Y_i^{(0)})]$, where $Y_i^{(0)}$ and $Y_i^{(0)}$ are respectively the outcomes under control and treatment conditions based on the underlying model. The approximated true superiority score is calculated from the true value of τ, which is $\gamma_{approx} \approx \frac{exp(\tau/\sqrt{2}))}{1+exp(\tau/\sqrt{2})}$. The averaged absolute bias of the estimated superiority score from the ordinal logistic regression is 0.035, while the estimated superiority score from the adjusted U-statistic is 0.013. The larger bias from the ordinal logistic regression may be due to the approximation procedure when we transform the treatment effect parameter τ to the superiority score γ. Since the U-statistic is essentially testing the superiority score, it has less bias than the OLR model.

3.3.2 Scenario 2: Outcome follows a mixture cumulative Box-Cox model

In practice, we usually do not have enough priori information and do not know the underlying model. In order to examine the performance of each method when the underlying model is unknown, we consider the simulation scenario 2 in which the simulated outcome variable is based on the mixture of Box-Cox distribution functions. Here X_i and T_i were generated as Scenario 1. However, the outcome variable Y_i was generated from a mixture distribution of the following form:

$$Y_i = T_i \cdot F_1(X_i, T_i) + (1 - T_i) \cdot F_0(X_i, T_i)$$

where $F_1(X,T) = F(\alpha_k + X\delta + \tau T; \lambda)$, $F_0(X,T) = F(\alpha_k + X\delta;; \lambda)$, and F belongs the Box-Cox family distributions (Guerrero and Johnson, 1982) with the following form:

$$F(x;\lambda) = \begin{cases} 0, & x < -\frac{1}{\lambda},\ \lambda > 0 \\ \frac{(1+\lambda x)^{1/\lambda}}{(1+\lambda x)^{1/\lambda}+1}, & 1+\lambda x > 0,\ \lambda \neq 0 \\ 1, & x > -\frac{1}{\lambda},\ \lambda < 0 \end{cases}$$

In the simulation study, we set $\lambda = 1$ and set $(\alpha_1, \alpha_2, \alpha_3)$ by calculating the probabilities of occurrence at the four levels of the outcome as (0.2, 0.1, 0.05, 0.65) for comparator group. The ordinal logistic regression model and the adjusted U-statistic are included in the power analysis. The superiority score estimated by the ordinal logistic regression model is approximated according to the transformation given by Agresti and Kateri (2016).

Figure 3.2 shows the power curve of the ordinal logistic regression and the adjusted U-statistic based on the simulated datasets from simulation Scenario 2. On the contrary of Scenario 1, the adjusted U-statistic has a higher power than the ordinal logistic regression when τ is small and the two methods are getting close when τ goes large. However, in this scenario the test size of the ordinal logistic regression model is 0.082 which is too liberal compared to 0.05. The simulation results in this scenario is reported in Table 3.3. The true superiority score is calculated as described in Section 3.3.1. The averaged absolute bias of the adjusted U-statistic is 0.012, which is much lower than the bias based on OLR model of 0.069. Compared with the bias in Scenario 1 where the underlying model is the OLR model, the bias of the OLR model significantly increases when the underlying model is misspecified, while the adjusted U-statistic keep a relatively small bias.

3.4 NHANES data application

For illustration, we applied the adjusted U-statistic and the ordinal logistic regression to estimate the treatment effect of physical activity on diabetic status with control the confounding variables such as the sociodemographic characters and dietary information based on the National Health and Nutrition Examination Survey (NHANES) datasets.

The NHANES is a nationally survey conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics (CDC-NCHS). NHANES collects both nutrition status and health conditions for the U.S. population. This study uses two NHANES datasets (2011-2012 and 2013-2014) in the analysis. The study population includes 7876 subjects from age 20 to 79 with complete 24-hour dietary recall data, complete laboratory data for plasma fasting glucose, glycohemoglobin, cholesterol and triglycerides, complete examination data for body measure and blood pressure, and complete questionnaire data for diabetes and physical activities.

The outcome variable in this study is defined as different status of diabetes and it is categorized to four levels: diabetes, pre-diabetes, metabolic syndrome and normal. Sociodemographic characteristics include gender, age, race, education and income. The dietary information of NHANES participants are collected by two 24-hour dietary recall interviews. The data of total nutrients intakes on the second time interviews are used in this study. 65 nutrients variables are included in the analysis. Physical activities is considered as the "treatment" variable and it is first categorized to three groups: no exercise, moderate exercise and vigorous exercise. However, we further categorize the physical activity into two leverls: subjects who have neither vigorous exercise nor moderate exercise are treated as no exercise; vigorous exercise and moderate exercise are combined to one group as having exercise in the analysis.

Principal component analysis is performed to summarize the dietary information and reduce the dimension of the dietary data. Components that had an eigenvalue of less than 1.0 (Kaiser, 1960) and accounted for less than 1% of variance were excluded. In the analysis, 65 variables were reduced to 13 principal components. Factor loadings with absolute values greater than 0.2 are considered. Each of the principal component is labeled by the nutrition variable with highest loadings. The 13 principal components are labeled as energy, beta carotene, long chain fatty acid, vitamin K, vitamin D, cholesterol, moisture, caffeine (negative), caffeine (positive), total sugar, short chain fatty acid, beta cryptoxanthin and alcohol.

Using the adjusted U statistics, the estimated superiority score is $\hat{\gamma}_u = 0.516$ ($p = 0.011$), which indicates that physical activity has significant protective effects on diabetes. The ordinal logistic regression reports the estimates of treatment parameter $\hat{\tau} = -0.156$ ($p = 0.010$), and the approximated superiority score is $\hat{\gamma}_{OLR} = 0.528$, which is similar to the results based on the adjusted U statistic.

The superiority measures $\hat{\gamma}$ has the interpretation that at any particular values for nutrition intakes, individuals with physical activities have approximately 52% chance (51.6% based on the adjusted U-statistic and 52.8% based on the OLR model) to have better diabetes status than the ones without physical activities.

3.5 Discussion and Conclusion

In this project, the causal parameter of the stochastic superiority score is considered to estimate the treatment effect for the ordinal outcomes. We investigated and compared the performance of different methods for estimating the treatment effect for the ordinal outcomes with control of the confounding covariates, including the ordinal logistic regression, the adjusted U-statistic, and the propensity score based stratification and matching. The power analyses in the simulation studies show that the adjusted U-statistic is more powerful and robust to estimate the superiority score

for the ordinal outcome.

The superiority score and its slight variation have also been studied under different experimental conditions for ordinal outcomes. For example, Chen (2009) considers to assess the treatment effect for the ordinal outcome when the noncompliance to assigned treatment exists. The procedure provides and estimates the form of the superiority score based on a function of the multinomial distributions of compliers to treatment assignments using likelihood method. Some other causal parameters are also proposed in the literatures for studying the treatment effect when the outcome is ordinal. Huang et al. (2017) proposes a plug-in estimator of the fraction who benefits when the outcome is ordinal based on the marginal distribution of the potential outcomes. The fraction who benefit is defined as the proportion of the population whose potential outcome in treatment group is better than the potential outcome in control groups, that is, $Pr(Y_T > Y_c)$, where Y_T and Y_c are the potential outcomes for treated subjects and untreated subjects, respectively. Lu et al. (2016) considers two causal parameters, the probabilities that the treatment is beneficial and strictly beneficial for the subjects. These two estimators can actually be applied for any outcomes but especially for the ordinal outcomes. Lu et al. (2016) provides the sharp bounds of the two parameters using the marginal distributions of the potential outcomes free from the joint distribution assumption of the potential outcomes.

The superiority of a particular method may depend on the data structure. When the data satisfies the proportional odds assumption, the ordinal logistic regression is no doubt a powerful way to estimate the treatment effect for the ordinal outcomes. However, when the model is misspecified, the ordinal logistic regression lacks robustness and may result in serious biased estimation. The adjusted U-statistics is a free of outcome-regression model and thus has higher robustness than the parametric approaches when the underlying model is unknown, which is the most commonly situation in practice. Since the U-statistic is adjusted by individual reweighting of the

data based on the propensity score, it is able to control the impact of the confounding covariates.

3.6 Tables and Figures

Table 3.1: Sizes of the tests based on different methods for simulated data sets generated under simulation Scenario 1.

Methods	Sizes
OLR	0.051
Matching	0.332
Stratification	0.262
Unadj. U Statistics	1.000
Adj. U Statistics	0.048

Table 3.2: Summarized simulation results for Scenario 1, where outcome was generated from an ordinal logistic regression model.

τ	True γ	Approximated true γ	Ordinal Logistic Regression			Adjusted U-Statistic		
			$\hat{\gamma}_{OLR}$	$\sigma_{\hat{\gamma}_{OLR}}$ *100	Empirical $\sigma_{\hat{\gamma}_{OLR}} * 100$	$\hat{\gamma}_u$	$\sigma_{\hat{\gamma}_u}$ *100	Empirical $\sigma_{\hat{\gamma}_u} * 100$
0.000	0.500	0.500	0.501	1.178	1.204	0.500	0.741	0.743
0.050	0.497	0.491	0.491	1.208	1.199	0.495	0.761	0.748
0.100	0.494	0.482	0.482	1.183	1.193	0.490	0.736	0.751
0.150	0.491	0.474	0.473	1.174	1.187	0.485	0.741	0.753
0.200	0.488	0.465	0.465	1.147	1.180	0.480	0.740	0.757
0.250	0.485	0.456	0.456	1.175	1.173	0.474	0.774	0.760
0.300	0.482	0.447	0.447	1.137	1.165	0.469	0.760	0.763
0.350	0.479	0.438	0.438	1.167	1.157	0.464	0.783	0.766
0.400	0.476	0.430	0.430	1.158	1.149	0.459	0.784	0.768
0.450	0.473	0.421	0.421	1.134	1.141	0.453	0.774	0.772
0.500	0.470	0.413	0.413	1.111	1.131	0.448	0.760	0.775
0.550	0.467	0.404	0.404	1.094	1.121	0.443	0.754	0.777
0.600	0.464	0.395	0.396	1.098	1.112	0.437	0.780	0.779

Table 3.3: Summarized simulation results for Scenario 2, where the outcome variable was generated from a mixture of Box-Cox distributions.

τ	True γ	Ordinal Logistic Regression			Adjusted U-Statistic		
		$\hat{\gamma}_{OLR}$	$\sigma_{\hat{\gamma}_{OLR}}$ *100	Empirical $\sigma_{\hat{\gamma}_{OLR}} * 100$	$\hat{\gamma}_u$	$\sigma_{\hat{\gamma}_u}$ *100	Empirical $\sigma_{\hat{\gamma}_u} * 100$
0.000	0.500	0.492	1.358	1.379	0.500	0.626	0.639
0.050	0.503	0.506	1.387	1.392	0.505	0.623	0.636
0.100	0.507	0.522	1.409	1.404	0.511	0.628	0.633
0.150	0.510	0.537	1.469	1.414	0.516	0.645	0.631
0.200	0.514	0.553	1.453	1.422	0.522	0.620	0.629
0.250	0.517	0.570	1.499	1.427	0.527	0.644	0.626
0.300	0.521	0.587	1.445	1.430	0.533	0.609	0.624
0.350	0.524	0.603	1.523	1.431	0.538	0.640	0.623
0.400	0.528	0.621	1.509	1.428	0.543	0.627	0.621
0.450	0.531	0.638	1.485	1.423	0.548	0.632	0.619
0.500	0.534	0.655	1.448	1.416	0.553	0.625	0.618
0.550	0.537	0.673	1.485	1.403	0.558	0.633	0.617
0.600	0.541	0.690	1.432	1.388	0.563	0.631	0.616

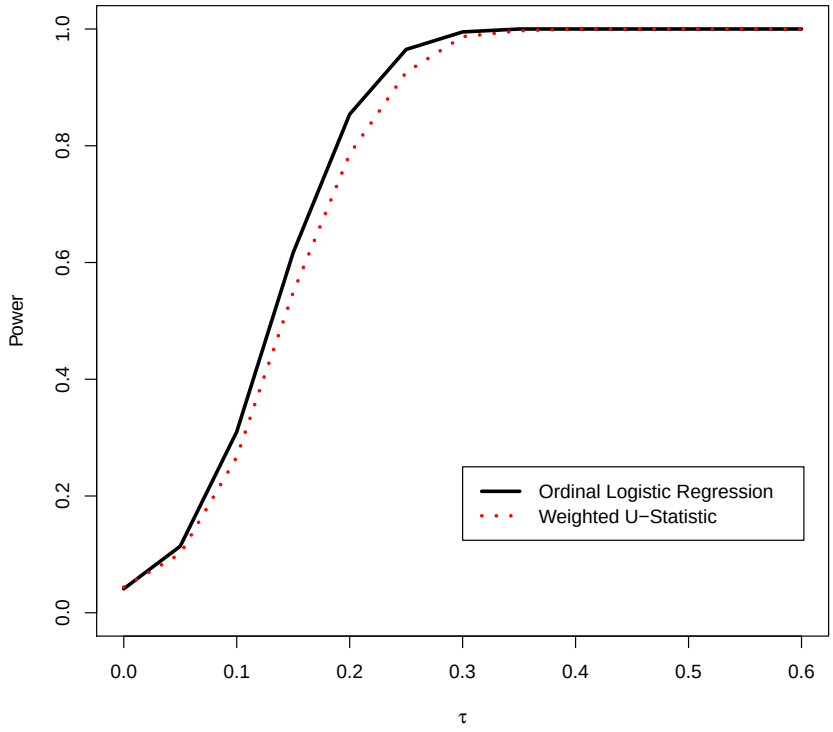

Figure 3.1: Power curves for different methods with data generated from ordinal logistic regression models in simulation Scenario 1.

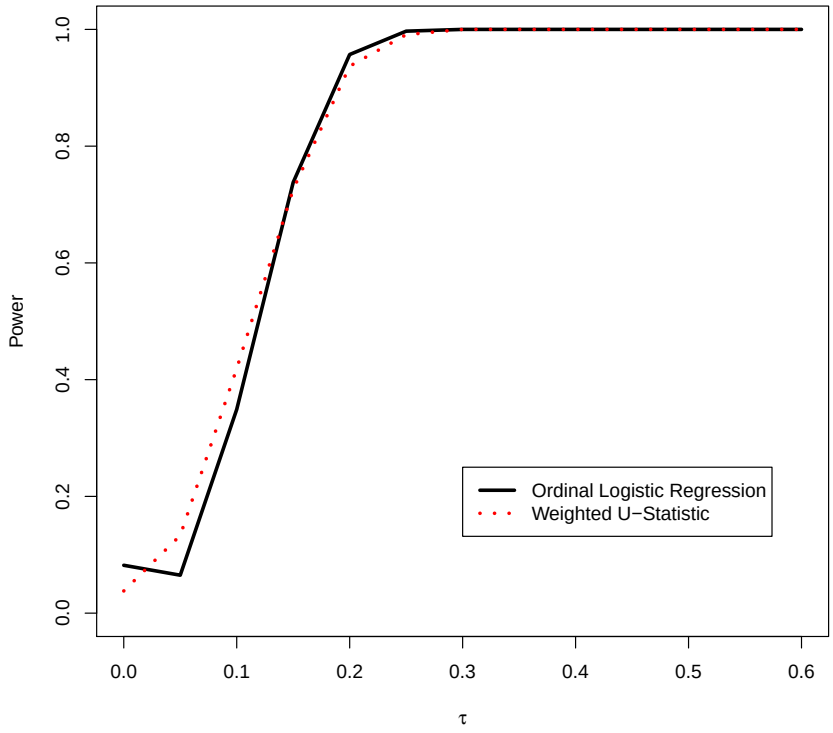

Figure 3.2: Power curves for different methods in with data generated under simulation Scenario 2.

CHAPTER 4

CAUSAL ANALYSIS OF DIETARY INFORMATION AND PHYSICAL ACTIVITY IN TYPE 2 DIABETES BY GENDER IN WHITE, AFRICAN AMERICAN AND MEXICAN AMERICAN: NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEYS 2011-2014

4.1 Introduction

Type 2 diabetes mellitus (T2DM) has become a major public health problem worldwide and one of the major causes of mortality in the United States outstripping cancer, HIV/AIDS, and cardiovascular disease. Enormous economic and psychosocial consequences are also associated with T2DM (Menke et al., 2015, Zimmet et al., 2016, American Diabetes Association, 2013 and Wang et al., 2016). The complications leaded by diabetes occur in many parts of body and increase the risk of ultimate death. Not only to the individual and their family, diabetes will also result in huge economic burden on the health system and nation. Globally, around 422 million adults were suffering from diabetes in 2014, which is 8.4% of the adult population (World Health Organization, 2016). In the U.S., the prevalence of diabetes has risen rapidly during the recent decades. In 1995, 3.3% of the U.S. population were diagnosed with diabetes and the number increased to 5.61% in 2005 and then 7.14% in 2015. In 2015, 30.3 million people in the U.S. have diabetes (9.4% of the U.S. population) of which 23.1 million has already been diagnosed as diabetes but

7.2 million has not been aware of or reported having diabetes (Centers for Disease Control and Prevention, 2017).

An analysis is performed by Menke et al. (2012) using the National Health and Nutrition Examination Survey (NHANES) data, which shows that the total unadjusted prevalence (diagnosed plus undiagnosed) of diabetes was 12.3% in the US population, and in 25.2% of those diabetes were undiagnosed (Menke et al., 2015). Differences of diabetes distributions were also observed in different ethnicity groups. The study of Menke et al. (2015) shows that the prevalence is much higher in non-Hispanic black (19.1%) and in Mexican American (14.5%) compared to the prevalence of total diabetes in non-Hispanic white (10.1%).

Among all the people with diabetes around the world, T2DM comprises approximately 90% of diabetic cases, and type 1 diabetes only accounts for about 10% of cases of diabetes (World Health Organization, 1999). However, type 1 diabetes is a different disease process characterized by total inability of the pancreas to produce insulin. It is important to analyze the relationship between diet and type 2 diabetes because food intake is the crucial variable in the control of T2DM. The second important component of T2DM control is physical activity. These two components are two of the three major foundations of diabetes therapies, the third being medication. The NHANES continuous database includes physical activity and detailed analysis of dietary components. These data have not been previously analyzed and may offer important insights into dietary variability, physical activity and the management of T2DM. Multivariate causal model analysis of the combination of the dietary data, measures of physical activity, BMI, across three major US ethnic groups may reveal variation in these variables that underlies major features in type 2 diabetes. Such information may contribute to a better understanding of type 2 diabetes variation among ethnic groups, and a better understanding of type 2 diabetes between female and male within each ethnic groups.

The analytical model is path analysis (i.e., causal structural equations) to explore the relationships between variables that are clearly causal and outcomes (e.g., type 2 diabetes). Path analysis proposed by Wright (1921) is a method used to determine whether a multivariate non-experimental dataset fits a particular causal model or not. Although path analysis is not applied to establish the causal relationship between variables, it is very powerful to examine and compare different complex hypothesis causal models and to find the most consistent one to the data (Streiner, 2005). As a result, the causal importance of independent variables on the outcomes can be quantitatively estimated.

In the present investigation, we apply path analysisi to examine the association between physical activity and diet on diabetes outcome using the NHANES data (2011-2014). Included T2DM patients are those who are taking only oral hypoglycemic agents (biguanides, sulfonylureas, thiazolidinedione-TZDs). Nelson et al. (2002) described diet and physical activity situations of U.S. adults with T2DM from NHANES III. However, analyses of the NHANES dietary data in a causal path model that includes physical activity and predicts HbA1c has not been conducted previously. The objective of this investigation is to analyze the causal paths that predict HbA1c using data on demographics, dietary, BMI, physical activity.

4.2 Method

4.2.1 Data and Materials

NHANES 2011-2014 datasets are analyzed in this study. NHANES is a nationally survey conducted by the National Center for Health Statistics in the Centers for Disease Control and Prevention (CDC-NCHS). It collects both of nutrition status and health conditions for the U.S. population. This project combines two NHANES datasets of study circles 2011-2012 and 2013-2014 in the analysis. The dietary information of

NHANES participants are collected by two 24-hour dietary recall interviews. The data of total nutrients intakes on the second time interviews are used in this study.

Totally, 65 nutrition intake variables are extracted from dietary dataset and used to measure dietary information of subjects in this project. Subjects from 20 to 79 years old in non-Hispanic white, non-Hispanic black and Mexican American are included in the study. Two dummy variables of non-Hispanic black and Mexican American are used to stand for the three different ethnicities in the study population. All the subjects included have a full record for demographic characteristics, physical activity, BMI, HbA1c and all 65 nutrition intakes.

The variable of physical activity is defined using the physical activity questionnaire data. Subjects who reports either vigorous work activities or vigorous recreational activities are defined as vigorous activity. Subjects who have moderate work activities or moderate recreational activities are defined as moderate activity. The final variable of physical activity used in the analysis is a dummy variable with 1 indicates subjects who have either vigorous activities or moderate activities.

Education includes five levels: less than 9th grade, 9-11th grade (including 12th grade with no diploma), high school graduate/general educational development or equivalent, some college or associate degree, and college graduate or above. Education is treated as a continuous variable by assigning scores 1 to 5 to the five levels in order. Higher score indicates higher education level.

Social economics status (SES) is measured by family monthly poverty level. It is categorized to 3 levels based on monthly poverty level index: ≤ 1.30, between 1.30 and 1.85, and > 1.85. The monthly poverty level index is defined as the ratio of monthly family income to the Department of Health and Human Services (HHS) poverty guidelines specific to family size. Numeric scores 1 to 3 are assigned to the three levels of SES. SES is treated as a continuous variable in the analysis where a higher score indicates a higher income.

4.2.2 Statistical Analysis

Path analysis is applied to determine the causal relationships between demographic variables, physical activity and dietary information to the outcome variable which is the diabetes status. Glycohemoglobin (HbA1c) level is used to measure the diabetes status. Figure 1 shows the hypothesis causal model considered in this study. In the assumed path diagram, Xs indicate the demographic variables or physical activity, and Ns indicate the dietary information. For illustration, only two demographic variables and two nutrition intake variables are drawn in Figure 4.1. The demographic variables or physical activities (Xs) are considered as exogenous variables, where the variance of these variables are not explained by the variables inside the assumed causal model. The dietary variables (Ns), BMI and HbA1c are the endogenous variables where part of their variances are explained by the other variables inside the causal model. In the path diagram, double headed arrow indicates the correlation between two variables which indicates that we assume there is no causal relationship between the two variables but there is correlation between them. The single headed arrow indicates the directional causal paths from one variable to another one. For all endogenous variables, part of their variance are also possibly explained by the extraneous variables (es), which are the variables outside the causal models.

In the hypothetical causal model, there are four different paths that the demographic variables or physical activity can influence the outcome: (1) direct effect: $X \rightarrow HbA1c$; (2) indirect effect through dietary information: $X \rightarrow N \rightarrow HbA1c$; (3) indirect effect through BMI: $X \rightarrow BMI \rightarrow HbA1c$; and (4) indirect effect through dietary information and BMI: $X \rightarrow N \rightarrow BMI \rightarrow HbA1c$. The total causal effects of demographic variables or physical activity on the outcome will be the sum of the effects from all the paths mentioned above.

All the data for continuous variables are standardized. A multivariate linear regression by regressing the outcome on all the other variables (demographic variables,

physical activity, dietary variables and BMI) is performed to identify the important variables in the path analysis. In the path analysis, each direct path is examined based on the Bonferroni adjusted p-values. Only significant paths are retained and reported in later path analysis results. An indirect paths is considered significant if the product of the path coefficients is greater than 0.1.

Path analysis is performed for the entire population first. The descriptive statistics of all the variables included in the path analysis for the entire population are reported in Table 4.1. In order to study the causality for each homogeneous group, path analysis is also performed for each gender-specific subpopulation and gender-ethnicity-specific subpopulation.

4.3 Results

2832 subjects are included in the analysis with age 48.1 $\pm$ 16.2 (mean$\pm$sd). 52.7% of the participants are female, and 47.3% are male. The study population includes three major ethnicities in the U.S.: 1573 (55.5%) non-Hispanic white, 814 (28.7%) non-Hispanic Black and 445 (15.7%) Mexican Americans. 7 dietary variables are significant from the multivariate linear regression of regressing HbA1c on all demographic variables, physical activity, 65 dietary variables and BMI. The 7 significant dietary variables are: protein, carbohydrate, total sugars, total fat, magnesium, moisture and butanoic fatty acid, which are included in the path analysis. Table 4.1 reports the descriptive statistics for the variables included in the path analysis by gender and race.

Figure 4.2 shows the path diagram with significant paths for the entire study population. Table 4.2 reports the estimated path coefficients of significant paths. From Table 4.2, physical activity does not show significant direct causal effect on HbA1c. However, physical activity indirectly affects HbA1c via BMI and magnesium intake. Via BMI, physical activity has a negative effect of -0.043 on HbA1c, and via

magnesium intake, the effect is -0.026. Physical activity thus has a negative causal effect of -0.069 on HbA1c in total.

Age positively affects HbA1c with a total causal effect of 0.298, which includes a direct effect of 0.282 and an indirect effect of 0.016 via BMI. Similarly, education affects HbA1c through a direct effect of -0.087 and an indirect effect of -0.019 via magnesium intake, which lead to a total causal effect of -0.106 on HbA1c. Social economics status does not show any significant causal effect on HbA1c.

For the 7 dietary variables included in the path analysis, butanoic fatty acid does not have significant causal relationship with HbA1c. Total sugars intake has a direct negative effect of -0.081 on HbA1c. Protein, carbohydrate, total fat and moisture all have an indirect effect on HbA1c through BMI, which are 0.030, -0.021, 0.024 and 0.026, respectively. Magnesium intake shows a relatively strong causal effect of -0.161 on HbA1c, which includes a direct effect of -0.127 and an indirect effect of -0.034 via BMI.

Gender plays an important role in the causal model. Although gender does not affect HbA1c directly, gender has several different paths that indirectly affects HbA1c through one or two mediate variables. The estimated effect of each path are reported in Table 2. In summery, gender has a total causal effect of -0.107 on HbA1c which indicates that female has higher HbA1c than male. In order to investigate the difference in the causal structure for between male and female, a separate path analysis is performed for male and female, respectively.

Table 4.3 and Table 4.4 report the estimated causal effects of each independent variable on the outcome variable HbA1c for male and female, respectively. Figure 4.3 and Figure 4.3 are the path diagram with only significant paths for male and female, respectively. Physical activity shows similar causal effects on HbA1c for both male (-0.057) and female (-0.051). However, for male, physical activity affects HbA1c via BMI, while for female, physical activity affects HbA1c through magnesium intake. For

female, age has a direct effect of 0.293 on HbA1c. For male, age has a direct effects of 0.245 on HbA1c, and also affects HbA1c via energy and carbohydrate intake with a total effect of 0.295. Education and SES do not show significant causal effects on HbA1c for male. However, SES has an indirect causal effect of -0.115 on HbA1c through BMI, which indicates that female with a lower income is believed to have a higher BMI, and then result in a higher HbA1c value.

Two dummy variables of Mexican American and Non-Hispanic black are used in the analysis and Non-Hispanic White is used as a reference. According to the results in Table 4.2, Mexican American shows a negative causal effect of -0.065 on HbA1c, which comes from the indirect paths via magnesium intake. Non-Hispanic black has a strong positive causal effect of 0.347 on HbA1c, which includes a direct effect of 0.302 and an indirect effect of 0.045 through BMI. In order to study the specific causal relationship for each homogeneous ethnicity group, separated path analysis are constructed for each gender-race-specific subpopulation.

Table 4.5 reports the significant paths for Mexican American male and female. Dietary variables in the path analysis for Mexican American male include total fat, added vitamin B12, vitamin K and caffeine. Dietary variables included in the path analysis for Mexican American female are: total sugars, total monounsaturated fatty acids, lutein and zeaxanthin, magnesium, iron and selenium. Age has strong positive direct effect of 0.385 for Mexican American male, compared to 0.322 for Mexican American female. Education and SES do not show any significant causal effects on HbA1c for Mexican American.

Table 4.6 reports the estimated causal effects of significant paths for non-Hispanic black by gender. Since no dietary variables are significant in the multivariate linear regression for non-Hispanic black male, the following path analysis is skipped. For non-Hispanic black female, physical activity shows a significant effect of -0.073 on HbA1c through magnesium intakes. Similarly, education also has a negative causal

effect of -0.052 on HbA1c through magnesium intakes. Age shows a strong positive direct effect of 0.340 on HbA1c for non-Hispanic black female. (See Table 4.6 for estimated path coefficients and Figure 4.6 for the path diagram).

Table 4.7 reports the estimates of significant path coefficients for non-Hispanic white by gender. For non-Hispanic white male, age has positive causal effect of 0.265 on HbA1c, which includes a direct effect of 0.227 and indirect effects through energy intake, carbohydrate intake and BMI. For non-Hispanic white female, age only has a direct effect of 0.273 on HbA1c. For white male, education shows a direct effect of -0.115 on HbA1c, while education is not significant for white female. However, physical activity causes reduction in BMI and hence results in an indirect effect on HbA1c of -0.081 for white female, while no significant effect of physical activity is observed for white male. Figures 4.7 and 4.8 are the path diagrams for non-Hispanic white male and female, respectively.

4.4 Discussion and Conclusion

In our study, the variation among different homogeneous groups are observed. Based on our analysis, males tend to have lower HbA1c level than females based on our analysis (-0.107, see Table 4.2). The total effect of physical activity on HbA1c is desired from several indirect paths through nutrition intake. Males have much higher intakes of all nutrition components than females, especially for protein (0.557) and total fat (0.447). Compared to males, females have a relatively higher BMI (-0.248), apparently causing a higher HbA1c. This result can be explained by the different body composition of female and male. The path model is also observed to vary across different ethnicities is also observed. Consistent with the prevalence of T2DM in the U.S. population, non-Hispanic blacks have a higher risk of T2DM compared to non-Hispanic whites. However, Mexican-Americans have a lower risk of T2DM than non-Hispanic white. Higher intake of magnesium in white, which is known to lead to

lower HbA1c values, is one of the significant differences.

Total sugar intake is observed to have a direct beneficial effect (-0.081, see Table 4.2) of decreasing HbA1c. Our finding supports the negative association between total sugar intake and the risk of T2DM observed from a prospective study of 36,787 non-diabetic men and women (Hodge et al., 2004). The observed effects of total sugar intake on T2DM from current published studies are inconsistent. Janket et al. (2003) reported that no definitive evidence was observed that total sugars intake is associated with developing T2DM based on a prospective study of 38,480 initially healthy postmenopausal women. Tsilas et al. (2017) performed a systematical review and meta analysis of prospective cohort studies to assess the association between the intakes of total sugar or certain sugar type and the risk of T2DM. No significant effect was observed of the intakes of total sugars and fructose on the risk of T2DM. However, intake of sucrose was observed to have beneficial effect of decreasing T2DM.

Higher intake of total fat is observed to result in higher HbA1c through BMI (0.024, see Table 4.2) in our study. This strong positive effect is particularly observed in males (0.342, see Table 4.3), while for females, total fat intake is not significant on HbA1c. Our result supports the conclusion of a prospective study of 42,504 male that total fat intake was associated with a higher risk of T2DM but was not independent of BMI in men (Van Dam et al., 2002). Also, no association was observed between total fat intake and the risk of T2DM in women based on another prospective study of 84,204 women (Salmeron et al., 2001). Although some studies reported no association between total fat intake and T2DM, significant relationship of the intakes of certain types of fat and T2DM were observed. Harding et al. (2004) conducted a prospective study of 23,631 Caucasian men and women aged 40-78 years, and there study didn't find significant association between total fat intake and the risk of T2DM. However, a higher ratio of dietary polyunsaturated fat to saturated fat was observed to have a protective effect on the development of T2DM.

In our study, intake of protein is found to have a positive effect on HbA1c through BMI (0.03, see Table 4.2). This result is consistent with some previous studies that intake of protein has positive relationship with the risk of T2DM (Malik et al., 2016; Sluijs et al., 2010; Van Nielen et al., 2014; Shang et al., 2016). Based on three large prospective cohort studies of US adults, Malik et al. (2016) also found a positive association between total protein intake and the risk of T2DM, and this positive association is observed to be largely due to the intake of animal protein. In the contrast, they also observed a protective effect of intake of vegetable protein on T2DM.

In our study, no direct effect of total carbohydrate intake on HbA1c is observed. However, dietary carbohydrate is negatively associated with BMI (-0.10), which further results in a lower effect (-0.021) on decreasing HbA1c (see Table 4.2). Our finding of the negative effect of carbohydrate intake on BMI is consistent with the literatures (Hodge et al., 2004). Merchant et al. (2009) reported that a low-carbohydrate diet is associated with higher probability of being overweight or obese among healthy adults. The association between carbohydrate intake and BMI or T2DM is controversial. Total dietary carbohydrate intake does not have significant effect on the risk of T2DM according to some researches. However, certain patterns of carbohydrate consumption (e.g., percentage of total energy) have a significant effect on BMI and HbA1c. Ma et al. (2005) reported no significant relationship between total daily carbohydrate intake and BMI among 572 healthy adults in central Massachusetts. However, the percentages of calories from carbohydrate had a significant positive effect on BMI. Meyer et al. (2000) reported that the total carbohydrate intake is associated with the risk of T2DM in a prospective cohort study of 35,988 Iowa women (55-69 years old) who were initially free of diabetes. However, certain types of carbohydrate, such as grains (especially whole grains), has significant protective effect of T2DM (Meyer et al., 2000). Haimoto et al. (2008) compared the long-term

effect of a carbohydrate-reduced diet to a conventional diet in T2DM patients in a 2-year follow up study. The conventional diet restricted energy intake by reducing fat for T2DM, while the carbohydrate-reduced diet asked T2DM patients limiting carbohydrate-rich foods to once or twice a day. They found that the carbohydrate-reduced diet caused decreases in both HbA1c and BMI. Their results suggested that restricting carbohydrate intake is more beneficial than restricting fat consumption in the management of T2DM.

Our study reveal the importance of magnesium intake in decreasing the risk of T2DM. Compared to the effect of protein, carbohydrate, total fat and moisture, magnesium intake has a stronger effect on HbA1c. Higher intake of magnesium has a better beneficial effect on decreasing BMI and HbA1c, further on decreasing the risk of T2DM. This finding is consistent with several other publications. Van Dam et al. (2006) found that high dietary magnesium intake is associated with a substantially lower risk of T2DM in U.S. black women in an 8-year prospective study of 41,186 U.S. black women. This result is consistent with ours that magnesium intake has a strong causal effect (-0.246) on decreasing HbA1c in non-Hispanic black females. The association between magnesium intake and fasting insulin concentrations are investigated among middle-aged women who were not diabetic (Fang et al., 2003). Higher magnesium intake is associated with lower fasting insulin concentrations, indicating higher insulin sensitivity. Lopez-Ridaura et al. (2004) observed a consistent inverse association between magnesium intake and risk of T2DM based on two large prospective studies in men and women. Across all the subgroup study, classified by different BMI, physical activity and family diabetes history, magnesium remained significant protection against T2DM. A meta-analysis of prospective cohort studies (Dong et al., 2011) supported the observations that magnesium intake has a significant inverse association with the risk of T2DM. Resnick et al. (1993) suggested that both extracellular and intracellular magnesium deficiency is associated with chronic mild

T2DM. A high prevalence of hypomagnesaemia and lower intracellular magnesium concentrations were found in diabetic subjects (Barbagal and Dominguez, 2006). Hypomagnesemia occurs among 13.5% to 47.7% among patients with T2DM compared with 2.5% to 15% among the subjects without T2DM (Pham et al., 2007). Magnesium has a complex relationship with insulin and glucose. On one hand, insulin is a hormone that regulates magnesium metabolism, and glucose is a physiologic determinant of cellular magnesium (Barbagal and Dominguez, 2007). On the other hand, magnesium is also a major determinant of insulin and glucose metabolism. It serves as a a modulator of insulin action and insulin sensitivity (Barbagal and Dominguez, 2007).

Another significant dietary component based on our analysis is moisture intake. It includes moisture present in all foods, beverages, and water consumed as a beverage. A positive association between moisture intake and HbA1c is observed through BMI (0.026, see Table 4.2). Our results supports the finding of Kant et al. (2009) that total water intake was associated with increased BMI. However, the association between total moisture intake and the risk of T2DM has not well-investigated because of the limited research, but the effects of various contributors to the total moisture intake on the risk of T2BM was examined in several studies. Roussel et al. (2011) reported a negative association between daily water intake and the risk of hyperglycemia based on a 9-year follow up study of 3,615 middle-aged French men and women. However, Pan et al. (2012) did not observe significant association between the plain water intake and the risk of T2DM based on a prospective study of 82,902 young and middle-aged women. In a cross-sectional study of sample size 138, Carroll et al. (2015) found that higher plain-water intake is associated with a lower T2DM risk score.

The relationship between BMI and T2DM has been well-investigated. BMI is an crucial and modifiable risk factor for T2DM. Nguyen et al.(2010) showed a clear association between BMI and diabetes in a large, representative sample of the

US population using NHANES 1999-2006 data. The lowest prevalence of diabetes was found in the normal weight group (BMI < 25.0), and the prevalence of diabetes increased as obesity class increases. Our study shows strong direct effect of BMI on increasing HbA1c across all subgroups. Some dietary intake factors(i.e., protein, carbohydrate, total fat and moisture) affect HbA1c only through BMI (see Table 4.2 and Figure 4.2). In order to have a clear understanding of causal paths through BMI, we also performed a path analysis that use two dummy variables (overweight with BMI between 25 and 30 and obesity with $BMI \geq 30$)) in the analysis instead of the continuous variable BMI. The results indicate that all the causal effects on HbA1c from independent variables through BMI are changed through obesity (see Figure 4.9 for the path diagram). These findings agree with the previously published results (Nguyen et al., 2010) that the prevalence of diabetes increases with increased obesity. The indirect effect from physical activity through obesity to HbA1c but not overweight also supports the finding that physical activity has a protective effect in subjects with high BMI (Hu, et al., 2004).

This project performed multivariate causal analysis to investigate the causal relationship between the two important components in T2DM therapies using the NHANES 2011-2014 continuous database: physical activity and dietary. Analysis was also carried out in homogeneous subgroups. Variability in causal structures are observed across different subgroups, while the causal importance of some components such as the intake of magnesium stay relatively stable. Magnesium intake is a significant protective factor, decreasing HbA1c, especially in females. Our work provides an insight understanding in possible causal association between physical activity, dietary components and diabetes status (i.e., HbA1c).

4.5 Tables and Figures

Table 4.1: Descriptive statistics for each variable in the analysis. Mean (SD) is reported for the continuous variable by gender and race. N (%) is reported for categorical variable.

Variables	Definition	Non-Hispanic White		Non-Hispanic Black		Mexican American	
		Male	Female	Male	Female	Male	Female
N	Sample Size	751	822	360	454	229	216
HbA1c	Glycohemoglobin (%). Blood glycohemoglobin is one of the measurements of assessing diabetes mellitus. The cut point of HbA1c is 6.5% for diagnosing diabetes.	5.6(0.9)	5.6(0.9)	6.1(1.6)	5.9(1.1)	5.8(1.2)	5.8(1.2)
BMI	Body Mass Index (kg/mˆ2). BMI=Weight (kg)/Height (m)ˆ2	29(6.3)	29.9(8.1)	29.4(6.7)	32.6(8.7)	29.3(5.3)	30.9(6.6)
Physical Activity	Subjects who have either vigorous or moderate work/recreational activities are considered as "having physical activities".	574(76.4%)	572(69.6%)	264(73.3%)	272(59.9%)	162(70.7%)	135(62.5%)
Age	Subjects from 20 to 79 years old are included in the analysis.	48.4(16.4)	49.1(16.5)	50.4(16.1)	46.8(15.7)	45.9(16.0)	45.2(14.7)
Education	Educations is considered as a continuous variable by assigning scores 1 to 5 to each level. A higher score indicates a higher education level.	3.7(1.1)	3.8(1.1)	3.4(1.1)	3.7(1.0)	2.6(1.3)	2.7(1.4)
SES	Social Econimics Satatus (SES) is treated as a continuous variable in the analysis where a higher score indicates a higher income.	2.3(1.1)	2.3(1.2)	2.2(1.2)	2.1(1.3)	1.9(1.3)	1.9(1.3)
Protein (gm)	Total daily protein intakes from foods, beverages and water	91.9(44.4)	68.3(31.7)	88.8(46.0)	69.4(38.8)	104.3(52.9)	75.6(37.5)
Carbohydrate (gm)	Total daily carbohydrate intakes from foods, beverages and water	277.2(130.2)	214.9(100.5)	247.1(123.4)	220.1(114.2)	299.9(144.6)	230.1(101.1)
Total sugars (gm)	Total daily sugars intakes from foods, beverages and water	123.5(83.5)	97.5(60.9)	107.2(70.9)	102.5(67.9)	118.6(67.9)	96.6(58.4)
Total fat (gm)	Total daily fat intakes from foods, beverages and water	88.3(45.3)	66.8(36.7)	80.9(46.6)	69.3(40.6)	90.1(52.4)	66.4(37.5)
Magnesium (mg)	Total daily magnesium intakes from foods, beverages and water	326.1168.6)	265.0(125.0)	288.7(159.5)	247.4(128.9)	367.1(196.7)	293.4(131.8)
Moisture (gm)	Total daily moisture intakes from foods, beverages and water	3120.9(1523.5)	2756.1(1260.8)	2552.3(1334.5)	2238.7(1182.2)	3099.7(1667.5)	2609.7(1152.3)
SFA 4:0 (Butanoic) (gm)	Total daily butanoic acid intakes from foods, beverages and water	0.6(0.6)	0.5(0.5)	0.4(0.4)	0.4(0.4)	0.5(0.6)	0.4(0.4)

Table 4.2: Significant paths and summarized causal effects for each variable in the path analysis for the entire population. Dietary variables in the path analysis include protein, total sugars, total fat, magnesium, moisture and butanoic fatty acid. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$".

Paths	**Intermediate Path Coefficients**			**Total Direct/Indirect Effects**
	r1	**r2**	**r3**	
Physical Activity				
Physical Activity ->BMI				-0.202
Physical Activity ->Magnesium				0.205
Physical Activity ->Moisture				0.184
Physical Activity ->BMI ->HbA1c		-0.202	0.214	-0.043
Physical Activity ->Magnesium ->HbA1c		0.205	-0.127	-0.026
Total effects on HbA1c				**-0.069**
Age				
Age ->HbA1c				0.282
Age ->BMI				0.074
Age ->Protein				-0.086
Age ->Carbohydrate				-0.124
Age ->Total_Sugars				-0.11
Age ->Total_Fat				-0.084
Age ->SFA_40_Butanoic				-0.085
Age ->BMI ->HbA1c		0.074	0.214	0.016
Total effects on HbA1c				**0.298**
Education				
Education ->HbA1c				-0.087
Education ->Protein				0.074
Education ->Magnesium				0.152
Education ->Moisture				0.078
Education ->Magnesium ->HbA1c		0.152	-0.127	-0.019
Total effects on HbA1c				**-0.106**
Gender				
Gender ->BMI				-0.248
Gender ->Protein				0.557
Gender ->Carbohydrate				0.442
Gender ->Total_Sugars				0.27
Gender ->Total_Fat				0.447
Gender ->Magnesium				0.373
Gender ->Moisture				0.261
Gender ->SFA_40_Butanoic				0.205
Gender ->BMI ->HbA1c		-0.248	0.214	-0.053
Gender ->Total_Sugars ->HbA1c		0.27	-0.081	-0.022
Gender ->Magnesium ->HbA1c		0.373	-0.127	-0.047
Gender ->Protein ->BMI ->HbA1c	0.557	0.139	0.214	0.017
Gender ->Total_Fat ->BMI ->HbA1c	0.447	0.111	0.214	0.011
Gender ->Magnesium ->BMI ->HbA1c	0.373	-0.161	0.214	-0.013
Total effects on HbA1c				**-0.107**
Mexican Americans				
MexA ->Protein				0.295
MexA ->Magnesium				0.401
MexA ->Magnesium ->HbA1c		0.401	-0.127	-0.051
MexA ->Magnesium ->BMI ->HbA1c	0.401	-0.161	0.214	-0.014
Total effects on HbA1c				**-0.065**
Non-Hispanic Black				
NHB ->HbA1c				0.302
NHB ->BMI				0.209
NHB ->Moisture				-0.359
NHB ->SFA_40_Butanoic				-0.266
NHB ->BMI ->HbA1c		0.209	0.214	0.045
Total effects on HbA1c				**0.347**
Total Sugars				
Total Sugars ->HbA1c				**-0.081**
Protein				
Protein ->BMI				0.139
Protein ->BMI ->HbA1c		0.139	0.214	0.030
Total effects on HbA1c				0.030
Carbohydrate				
Carbohydrate ->BMI				-0.1
Carbohydrate ->BMI ->HbA1c		-0.100	0.214	-0.021
Total effects on HbA1c				**-0.021**
Total Fat				
Total Fat ->BMI				0.111
Total Fat ->BMI ->HbA1c		0.111	0.214	0.024
Total effects on HbA1c				**0.024**
Magnesium				
Magnesium ->HbA1c				-0.127
Magnesium ->BMI				-0.161
Magnesium ->BMI ->HbA1c		-0.161	0.214	-0.034
Total effects on HbA1c				**-0.161**
Moisture				
Moisture ->BMI				0.121
Moisture ->BMI ->HbA1c		0.121	0.214	0.026
Total effects on HbA1c				**0.026**
BMI				
BMI ->HbA1c				**0.214**

Table 4.3: Significant paths and summarized causal effects for each variable in the path analysis for male. Dietary variables in the path analysis include energy, protein, carbohydrate, total sugars, total fat and moisture. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$".

Paths	**Intermediate Path Coefficients**			**Total Direct/Indirect Effects**
	r1	**r2**	**r3**	
Physical Activity				
Physical Activity ->BMI				-0.224
Physical Activity ->Moisture				0.239
Physical Activity ->BMI ->HbA1c		-0.224	0.256	-0.057
Total effects on HbA1c				**-0.057**
Age				
Age ->HbA1c				0.245
Age ->Energy				-0.136
Age ->Carbohydrate				-0.135
Age ->Total Sugars				-0.121
Age ->Energy ->HbA1c		-0.136	-0.473	0.064
Age ->Carbohydrate ->HbA1c		-0.135	0.223	-0.03
Age ->Energy ->BMI ->HbA1c	-0.136	-0.452	0.256	0.016
Total effects on HbA1c				**0.295**
Mexican American				
MexA ->Protein				0.321
MexA ->Protein ->BMI ->HbA1c	0.321	0.117	0.256	0.01
Non-Hispanic Black				
NHB ->HbA1c				0.378
NHB ->Carbohydrate				-0.239
NHB ->Total_Sugars				-0.233
NHB ->Moisture				-0.375
NHB ->Carbohydrate ->HbA1c		-0.239	0.223	-0.053
NHB ->Moisture ->BMI ->HbA1c	-0.375	0.107	0.256	-0.01
Total effects on HbA1c				**0.315**
Energy				
Energy ->HbA1c				-0.473
Energy ->BMI				-0.452
Energy ->BMI ->HbA1c		-0.452	0.256	-0.116
Total effects on HbA1c				**-0.589**
Protein				
Protein ->BMI				0.117
Protein ->BMI ->HbA1c		0.117	0.256	**0.030**
Carbohydrate				
Carbohydrate ->HbA1c				0.223
Carbohydrate ->BMI				0.075
Carbohydrate ->BMI ->HbA1c		0.075	0.256	0.019
Total effects on HbA1c				**0.242**
Total Fat				
Total Fat ->HbA1c				0.271
Total Fat ->BMI				0.277
Total Fat ->BMI ->HbA1c		0.277	0.256	0.071
Total effects on HbA1c				**0.342**
Moisture				
Moisture ->BMI				0.107
Moisture ->BMI ->HbA1c		0.107	0.256	**0.027**
BMI				
BMI ->HbA1c				**0.256**

Table 4.4: Significant paths and summarized causal effects for each variable in the path analysis for female. Dietary variables in the path analysis include energy, protein, carbohydrate, total sugars, total fat and moisture. Only significant direct effects and indirect effects no less than 0.01 are reported in the table. For indirect effects, intermediate path coefficients are reported in 2^{nd} to 4^{th} columns under the column title "$r1, r2, r3$".

Paths	**Intermediate Path Coefficients**			**Total Direct/Indirect Effects**
	r1	**r2**	**r3**	
Physical Activity				
Physical Activity ->Magnesium				0.230
Physical Activity ->Magnesium ->HbA1c		0.230	-0.171	-0.039
Physical Activity ->Magnesium ->BMI ->HbA1c	0.230	-0.257	0.197	-0.012
Total effects on HbA1c				**-0.051**
Age				
Age ->HbA1c				**0.293**
Mexican American				
MexA ->Phosphorus				0.244
MexA ->Magnesium				0.389
MexA ->Magnesium ->HbA1c		0.389	-0.171	-0.067
MexA ->Phosphorus ->BMI ->HbA1c	0.244	0.214	0.197	0.010
MexA ->Magnesium ->BMI ->HbA1c	0.389	-0.257	0.197	-0.020
Total effects on HbA1c				**-0.077**
Non-Hispanic Black				
NHB ->HbA1c				0.273
NHB ->BMI				0.382
NHB ->Moisture				-0.338
NHB ->BMI ->HbA1c		0.382	0.197	0.075
Total effects on HbA1c				**0.348**
SES				
SES ->BMI				-0.096
SES ->BMI ->HbA1c		-0.096	0.197	-0.019
Total effects on HbA1c				**-0.115**
Phosphorus				
Phosphorus ->BMI				0.214
Phosphorus ->BMI ->HbA1c		0.214	0.197	**0.042**
Magnesium				
Magnesium ->HbA1c				-0.171
Magnesium ->BMI				-0.257
Magnesium ->BMI ->HbA1c		-0.257	0.197	-0.050
Total effects on HbA1c				**-0.478**
Moisture				
Moisture ->BMI				0.134
Moisture ->BMI ->HbA1c		0.134	0.197	**0.026**
BMI				
BMI ->HbA1c				**0.197**

Table 4.5: Significant paths and summarized causal effects for each variable in the path analysis for Mexican American by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table.

Paths	**Estimates**
Mexican American Male	
Age ->HbA1c	0.385
Mexican American Female	
Age ->HbA1c	0.322
Magnesium ->HbA1c	-0.307
Magnesium ->BMI	-0.357

Table 4.6: Significant paths and summarized causal effects for each variable in the path analysis for non-Hispanic black by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table.

Paths	Intermediate Path Coefficients		Total Direct/Indirect Effects
	r1	r2	
Non-Hispanic Black Male			
	-	-	-
Non-Hispanic Black Female			
Age			
Age ->HbA1c			**0.340**
Physical Activity			
Physical Activity ->Magnesium			0.295
Physical Activity ->Magnesium ->HbA1c	0.295	-0.246	**-0.073**
Education			
Education ->Magnesium			0.212
Education ->Magnesium ->HbA1c	0.212	-0.246	**-0.052**
Energy			
Energy ->BMI			0.328
Energy ->BMI ->HbA1c	0.328	0.145	**0.048**
Carbohydrate			
Carbohydrate ->BMI			-0.489
Carbohydrate ->BMI ->HbA1c	-0.489	0.145	**-0.071**
Phosphorus			
Phosphorus ->HbA1c			0.297
Phosphorus ->BMI			0.312
Phosphorus ->BMI ->HbA1c	0.312	0.145	**0.045**
Magnesium			
Magnesium ->HbA1c			**-0.246**
BMI			
BMI ->HbA1c			**0.145**

Table 4.7: Significant paths and summarized causal effects for each variable in the path analysis for non-Hispanic white by gender. Only significant direct effects and indirect effects no less than 0.01 are reported in the table.

Paths	Intermediate Path Coefficients			Total Direct/Indirect Effects
	r1	r2	r3	
Non-Hispanic White Male				
Age				
Age ->HbA1c				0.227
Age ->Energy				-0.149
Age ->Carbohydrate				-0.168
Age ->Total_Sugars				-0.167
Age ->Energy ->HbA1c		-0.149	-0.578	0.086
Age ->Energy ->BMI ->HbA1c	-0.149	0.286	0.239	-0.01
Age ->Carbohydrate ->HbA1c		-0.168	0.322	-0.054
Age ->Carbohydrate ->BMI ->HbA1c	-0.168	-0.39	0.239	0.016
Total effects on HbA1c				**0.265**
Education				
Education ->HbA1c				**-0.115**
Energy				
Energy ->HbA1c				-0.578
Energy ->BMI				0.286
Energy ->BMI ->HbA1c		0.286	0.239	0.069
Total effects on HbA1c				**-0.509**
Protein				
Protein ->HbA1c				**0.143**
Carbohydrate				
Carbohydrate ->HbA1c				0.322
Carbohydrate ->BMI				-0.39
Carbohydrate ->BMI ->HbA1c		-0.39	0.239	**-0.093**
Total Fat				
Total Fat ->HbA1c				**0.212**
Thiamin (Vitamin B1)				
Thiamin VitB1 ->BMI				0.122
Thiamin VitB1 ->BMI ->HbA1c		0.122	0.239	**0.029**
BMI				
BMI ->HbA1c				**0.239**
Non-Hispanic White Female				
Age				
Age ->HbA1c				**0.273**
Physical Activity				
Physical Activity ->BMI				-0.356
Physical Activity ->BMI ->HbA1c		-0.356	0.228	**-0.081**
Education				
Education ->VitC				0.170
Education ->Alcohol				0.119
Alcohol				
Alcohol ->BMI				-0.193
Alcohol ->BMI ->HbA1c		-0.193	0.228	**-0.044**
SFA 4:0 (Butanoic)				
SFA_40_Butanoic ->BMI				-0.136
SFA_40_Butanoic ->BMI ->HbA1c		-0.136	0.228	**-0.031**
BMI				
BMI ->HbA1c				**0.228**

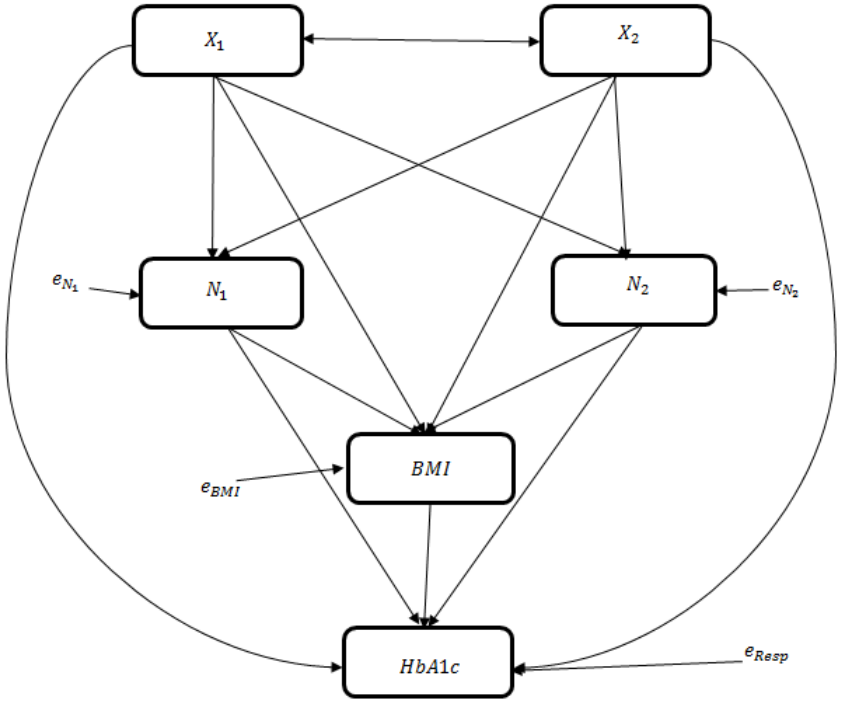

Figure 4.1: Hypothesis causal model (Adapted from Wright, 1934). X_1 and X_2 indicate demographic characteristics or physical activity. N_1 and N_2 are nutrition intake variables, which reflect the dietary information.

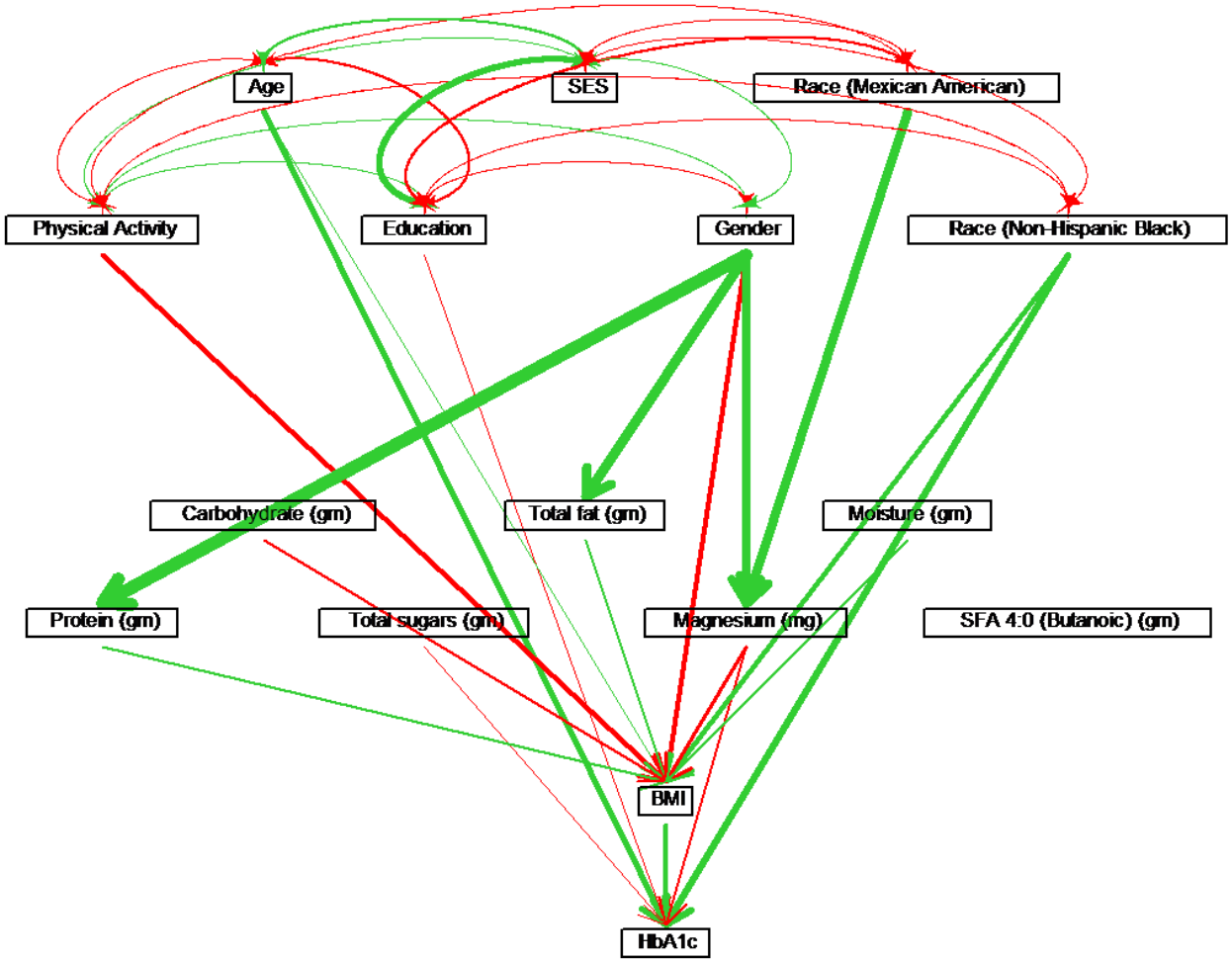

Figure 4.2: Path diagram for the entire sample. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

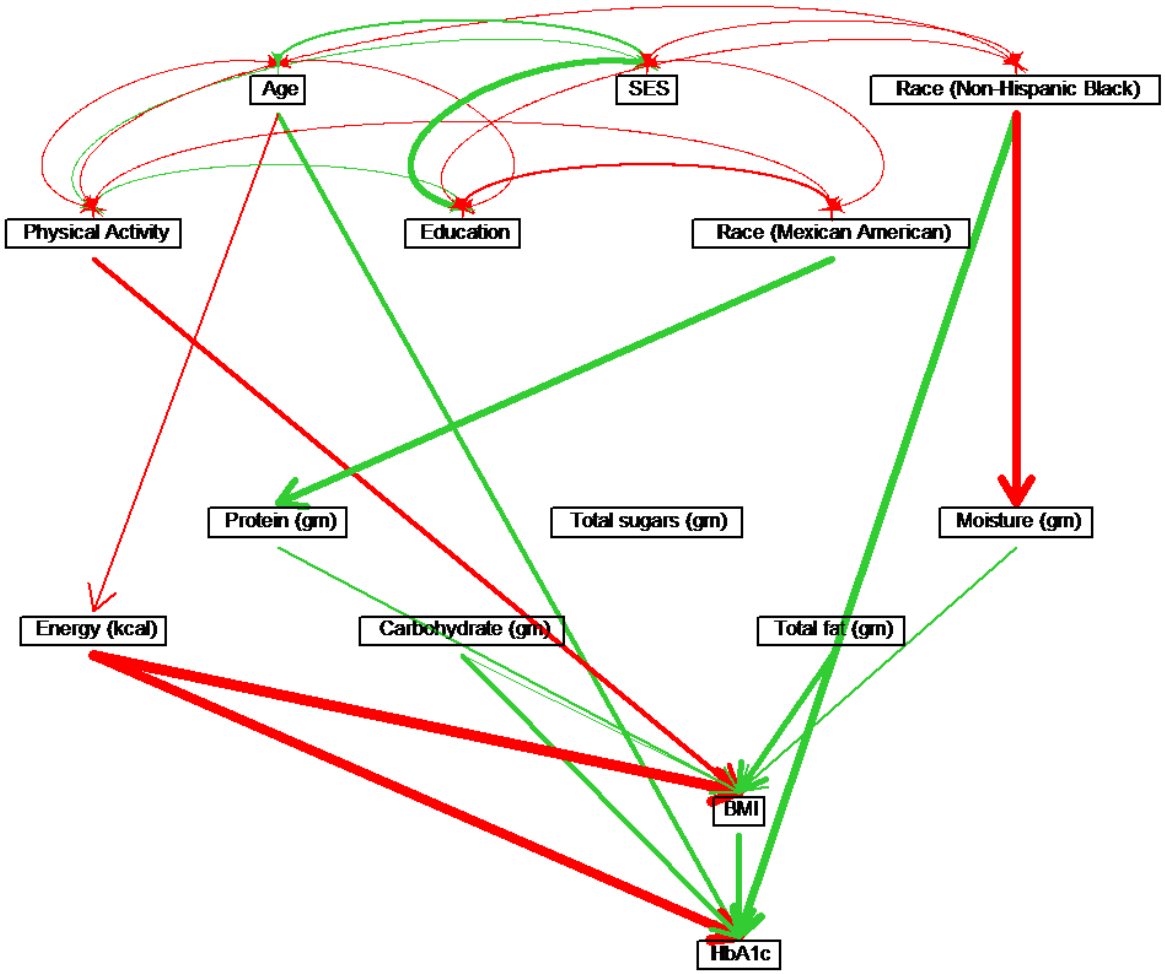

Figure 4.3: Path diagram for male. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients,where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

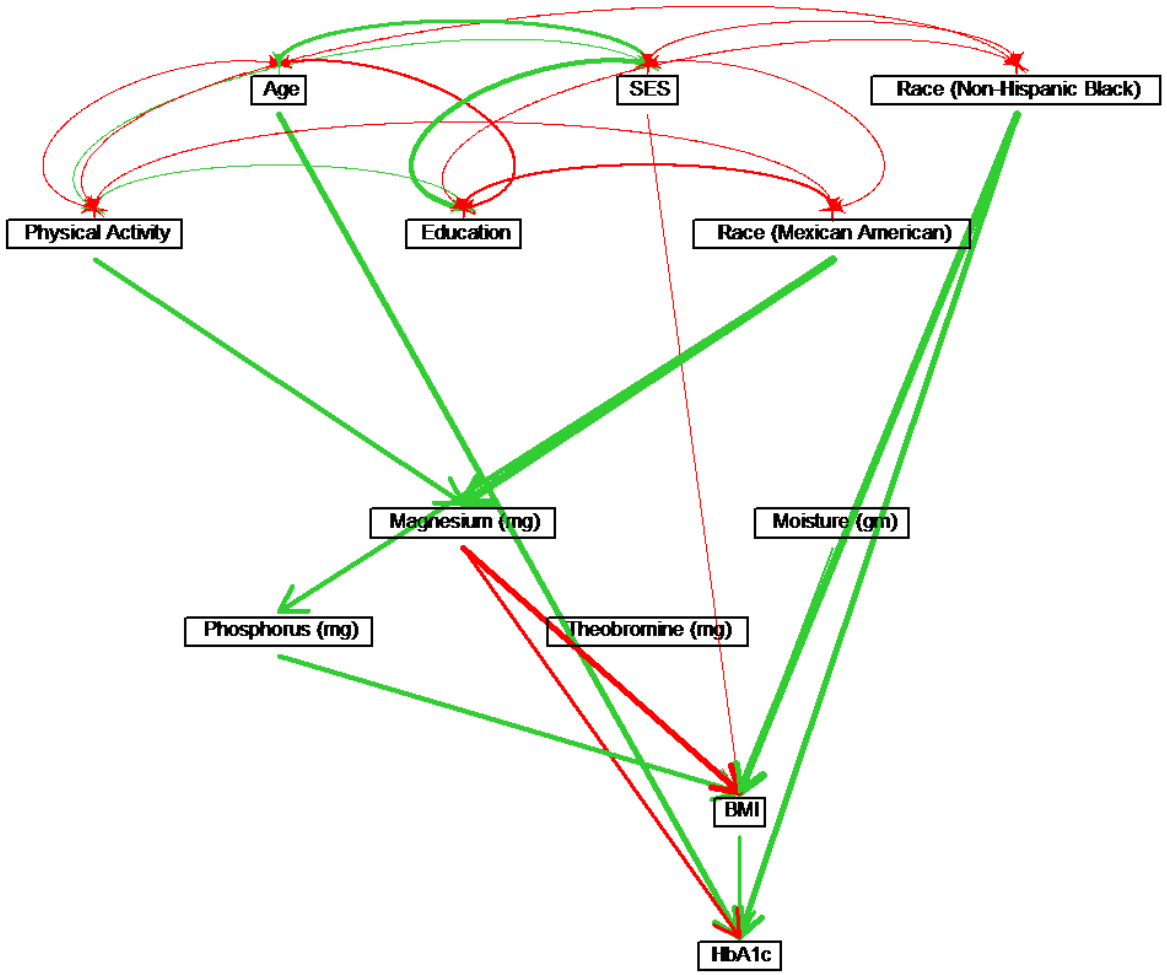

Figure 4.4: Path diagram for female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

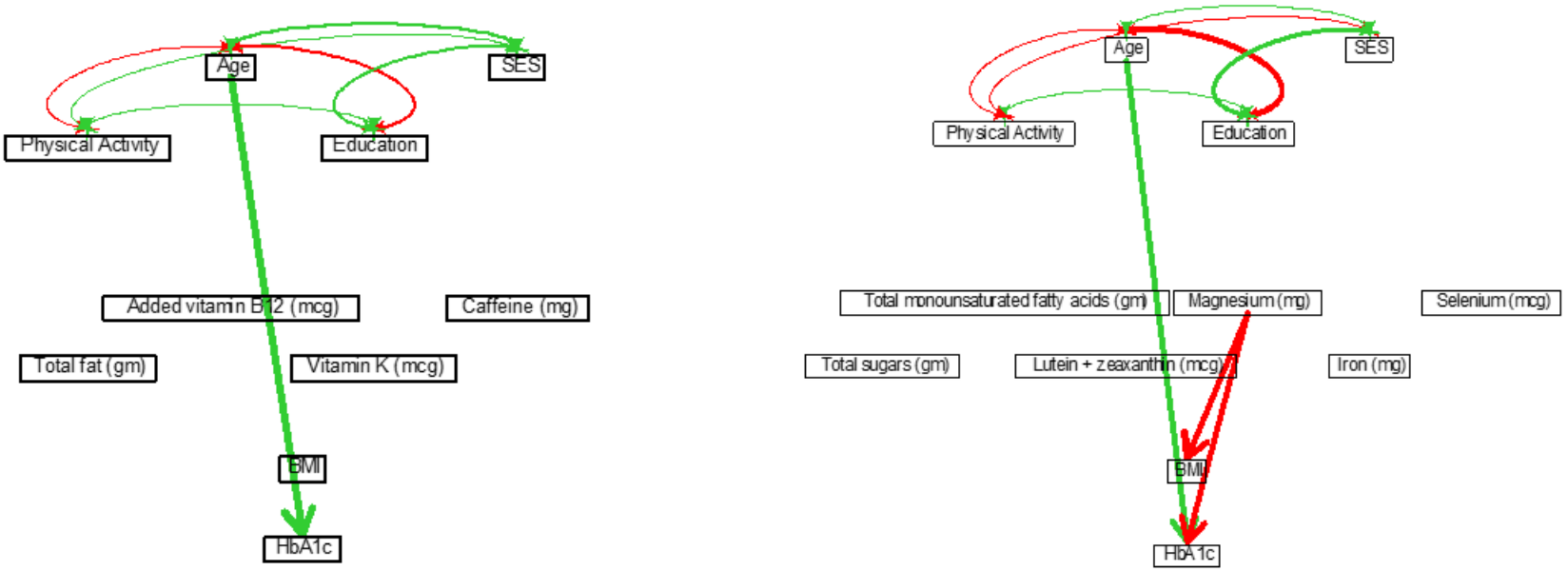

Figure 4.5: Path diagrams for Mexican American. Left panel is the diagram for Mexican American male and right one is for Mexican American female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

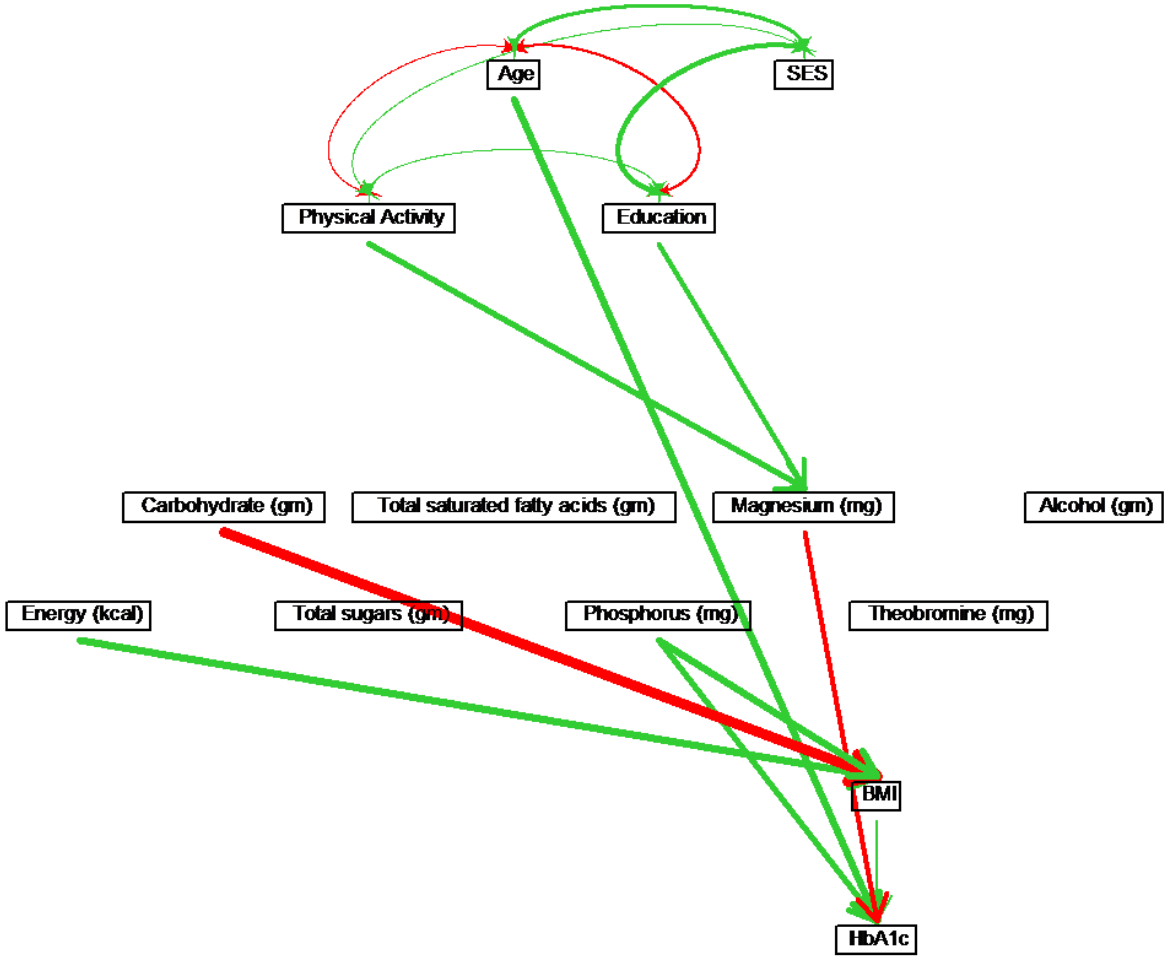

Figure 4.6: Path diagrams for non-Hispnic black female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

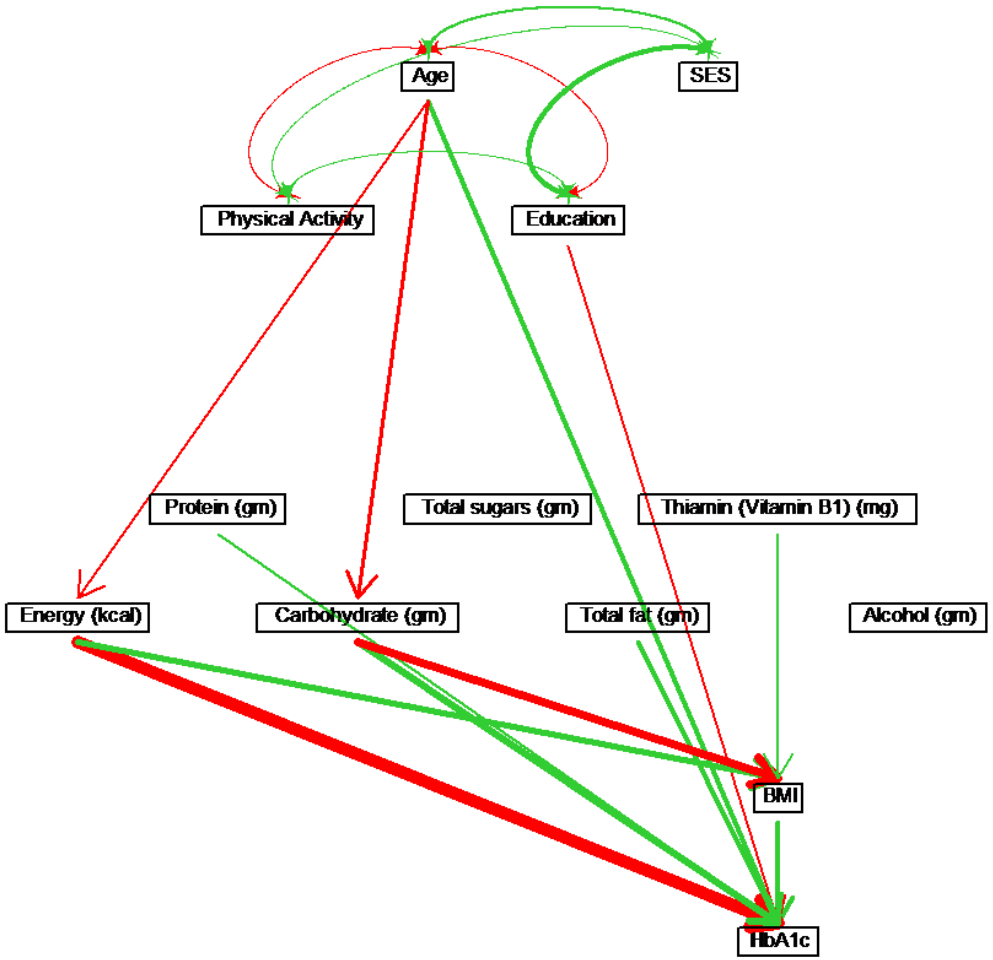

Figure 4.7: Path diagrams for non-Hispnic white male. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

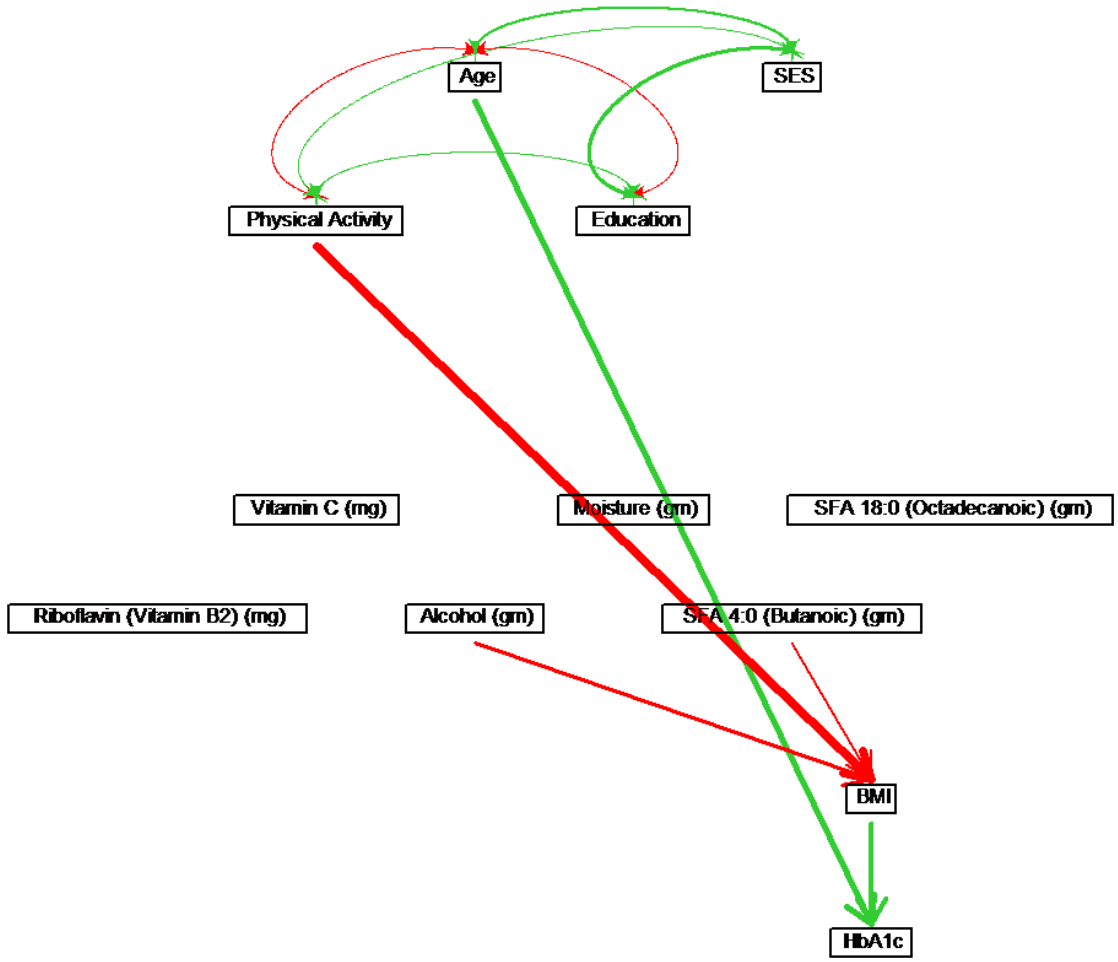

Figure 4.8: Path diagrams for non-Hispnic white female. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients,where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

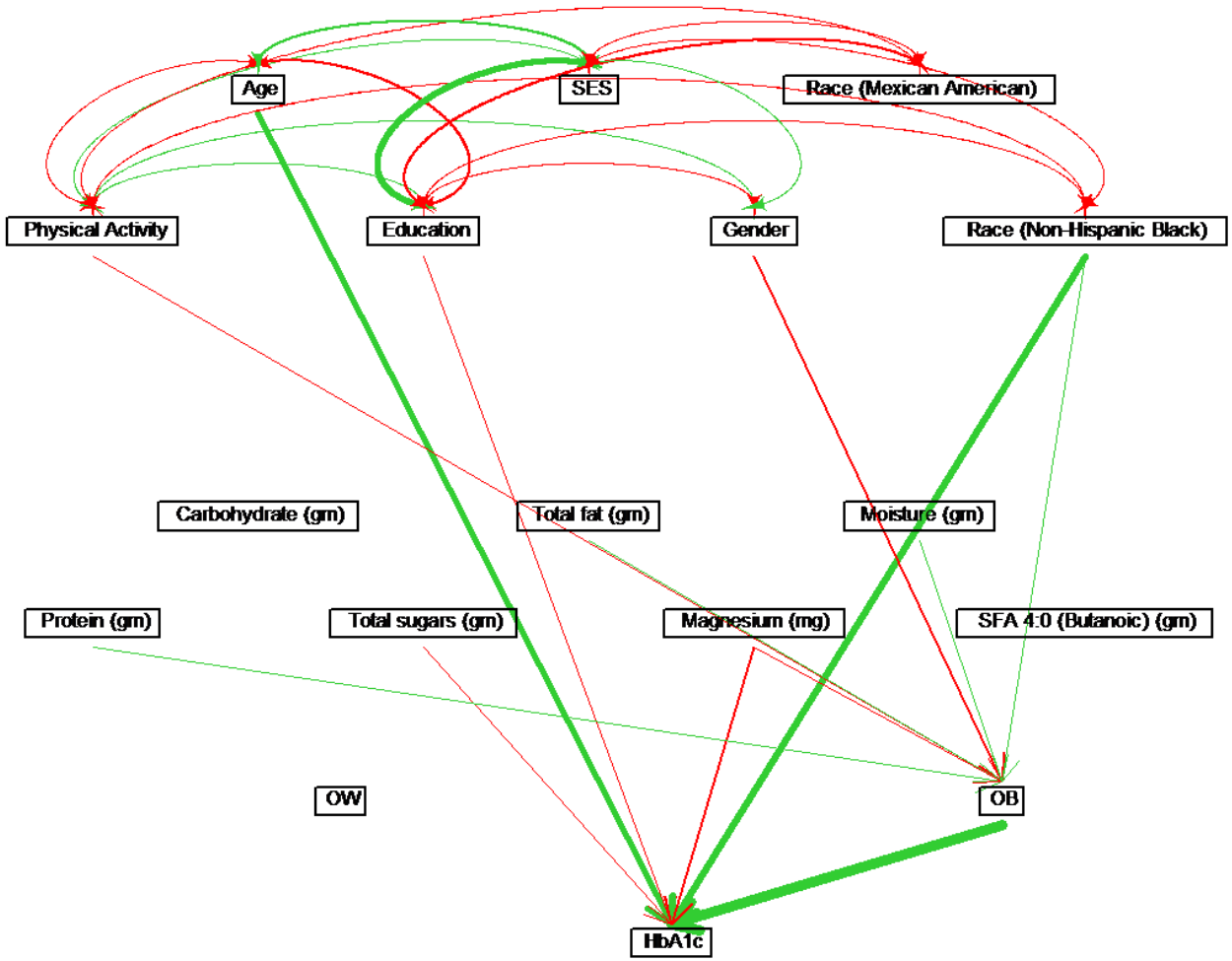

Figure 4.9: Path diagram for whole population using two dummy variables to measure BMI. Green paths indicate positive path coefficients, while red paths indicate negative coefficients. The width of the paths are related to the absolute values of path coefficients, where higher absolute value (i.e., wider paths) indicates stronger causality. Only significant paths are shown in the diagram.

REFERENCES

[1] Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. The Annals of Statistics, 32(3), 870-897.

[2] Berger, A. H., Niki, M., Morotti, A., Taylor, B. S., Socci, N. D., Viale, A., Taylor, B.S., Socci, N.D., Viale, A., Brennan, C., Szoke, J., Motoi, N., Rothman, P.B. & Teruya-Feldstein, J. (2010). Identification of DOK genes as lung tumor suppressors. Nature Genetics, 42(3), 216-223.

[3] Bonnas, C., Specht, K., Spleiss, O., Froehner, S., Dietmann, G., Krüger, J. M., Arbogast, S. & Feuerhake, F. (2012). Effects of cold ischemia and inflammatory tumor microenvironment on detection of PI3K/AKT and MAPK pathway activation patterns in clinical cancer samples. International Journal of Cancer, 131(7), 1621-1632.

[4] Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2), 465-480.

[5] Chen, L., Du-Cuny, L., Moses, S., Dumas, S., Song, Z., Rezaeian, A. H., Lin, H.K., Meuillet, E.J. & Zhang, S. (2015). Novel inhibitors induce large conformational changes of GAB1 pleckstrin homology domain and kill breast cancer cells. PLoS Computational Biology, 11(1), e1004021.

[6] Espina, V., Edmiston, K. H., Heiby, M., Pierobon, M., Sciro, M., Merritt, B., Banks, S., Deng, J., VanMeter, A. J., Geho, D. H. & Pastore, L. (2008). A por-

trait of tissue phosphoprotein stability in the clinical tissue procurement process. Molecular & Cellular Proteomics, 7(10), 1998-2018.

[7] Gajadhar, A. S., Johnson, H., Slebos, R. J., Shaddox, K., Wiles, K., Washington, M. K., Herline, A. J., Levine, D. A., Liebler, D. C. & White, F. M. (2015). Phosphotyrosine signaling analysis in human tumors is confounded by systemic ischemia-driven artifacts and intra-specimen heterogeneity. Cancer Research, 75(7), 1495-1503.

[8] George, E. I.,& McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889.

[9] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, 1(3), 515-534.

[10] Gu, H., & Neel, B. G. (2003). The Gabin signal transduction. Trends in Cell Biology, 13(3), 122-130.

[11] Gündisch, S., Grundner-Culemann, K., Wolff, C., Schott, C., Reischauer, B., Machatti, M., Groelz, D., Schaab, C., Tebbe, A. & Becker, K. F. (2013). Delayed times to tissue fixation result in unpredictable global phosphoproteome changes. Journal of Proteome Research, 12(10), 4424-4434.

[12] Hunter, T. (2009). Tyrosine phosphorylation: thirty years and counting. Current Opinion in Cell Biology, 21(2), 140-146.

[13] Hwang, J. G., & Peddada, S. D. (1994). Confidence interval estimation subject to order restrictions. The Annals of Statistics, 67-93.

[14] Lee, S., Huang, H., Niu, Y., Tommasino, M., Lenoir, G., & Sylla, B. S. (2007). Dok1 expression and mutation in Burkitt's lymphoma cell lines. Cancer Letters, 245(1), 44-50.

[15] Mertins, P., Yang, F., Liu, T., Mani, D. R., Petyuk, V. A., Gillette, M. A., Clauser, K. R., Qiao, J. W., Gritsenko, M. A., Moore, R. J. & Levine, D. A. (2014). Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. Molecular & Cellular Proteomics, 13(7), 1690-1704.

[16] Montana, G., Berk, M., & Ebbels, T. (2011). Modelling short time series in metabolomics: a functional data analysis approach. In Software Tools and Algorithms for Biological Systems (pp. 307-315). Springer New York.

[17] Némorin, J. G., Laporte, P., Brub, G., & Duplay, P. (2001). p62dok negatively regulates CD2 signaling in Jurkat cells. The Journal of Immunology, 166(7), 4408-4415.

[18] Paul, M. K., & Mukhopadhyay, A. K. (2004). Tyrosine kinaserole and significance in cancer. International Journal of Medical Sciences, 1(2), 101.

[19] Peddada, S. D., Lobenhofer, E. K., Li, L., Afshari, C. A., Weinberg, C. R., & Umbach, D. M. (2003). Gene selection and clustering for time-course and doseresponse microarray experiments using order-restricted inference. Bioinformatics, 19(7), 834-841.

[20] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[21] Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics, 38(5), 2587-2619.

[22] Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1999). WinBUGS Version 1.2 User Manual, MRC Biostatistics Unit.

[23] Tandon, M., Vemula, S. V., & Mittal, S. K. (2011). Emerging strategies for EphA2 receptor targeting for cancer therapeutics. Expert Opinion on Therapeutic Targets, 15(1), 31-51.

[24] Wu, H., Yuan, M., Kaech, S. M., & Halloran, M. E. (2007). A statistical analysis of memory CD8 T cell differentiation: An application of a hierarchical state space model to a short time course microarray experiment. The Annals of Applied Statistics, 442-458.

[25] Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. Biometrical Journal.

[26] Agresti, A. (2007). An introduction to categorical data analysis. John Wiley.

[27] Agresti, A., & Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. Biometrics, 73(1), 214-219.

[28] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399-424.

[29] Boes, S. (2013). Nonparametric analysis of treatment effects in ordered response models. Empirical Economics, 44(1), 81-109.

[30] Cheng, J. (2009). Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. Biometrics, 65(1), 96-103.

[31] Glen, A., Kong, M. & Datta, S. (2017). Multi-sample adjusted u-statistics that account for confounding covariates.

[32] Grissom, R. J. (1994). Statistical analysis of ordinal categorical status after therapies. Journal of Consulting and Clinical Psychology, 62(2), 281.

[33] Guerrero, V. M., & Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. Biometrika, 69(2), 309-314. Chicago

[34] Huang, E. J., Fang, E. X., Hanley, D. F., & Rosenblum, M. (2017). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few?. Biostatistics, 18(2), 308-324.

[35] Janket, S. J., Manson, J. E., Sesso, H., Buring, J. E., & Liu, S. (2003). A prospective study of sugar intake and risk of type 2 diabetes in women. Diabetes Care, 26(4), 1008-1015.

[36] Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 141-151. Chicago

[37] Klotz, J. H. (1966). The Wilcoxon, ties, and the computer. Journal of the American Statistical Association, 61(315), 772-787.

[38] Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology.

[39] Lu, J., Ding, P., & Dasgupta, T. (2015). Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds. arXiv:1507.01542.

[40] Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 50-60.

[41] McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Statistics in Medicine, 32(19), 3388-3414.

[42] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

[43] Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1), 33-38.

[44] Ryu, E., & Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. Statistics in Medicine, 27(10), 1703-1717.

[45] Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics, 23(2), 170-192.

[46] Whegang, S. Y., Basco, L. K., Gwt, H., & Thalabard, J. C. (2010). Analysis of an ordinal outcome in a multicentric randomized controlled trial: application to a 3-arm anti-malarial drug trial in Cameroon. BMC Medical Research Methodology, 10(1), 58.

[47] American Diabetes Association (2013). Economic costs of diabetes in the US in 2012. Diabetes Care, 36(4), 1033-1046.

[48] Austin, G. L., Ogden, L. G., & Hill, J. O. (2011). Trends in carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight,

and obese individuals: 19712006. The American journal of clinical nutrition, 93(4), 836-843.

[49] Barbagallo, M., & Dominguez, L. J. (2007). Magnesium metabolism in type 2 diabetes mellitus, metabolic syndrome and insulin resistance. Archives of Biochemistry and Biophysics, 458(1), 40-47.

[50] Carroll, H. A., Davis, M. G., & Papadaki, A. (2015). Higher plain water intake is associated with lower type 2 diabetes risk: a cross-sectional study in humans. Nutrition Research, 35(10), 865-872.

[51] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2017. https://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html.

[52] Fung, T. T., Manson, J. E., Solomon, C. G., Liu, S., Willett, W. C., & Hu, F. B. (2003). The association between magnesium intake and fasting insulin concentration in healthy middle-aged women. Journal of the American College of Nutrition, 22(6), 533-538.

[53] Gaesser, G. A. (2007). Carbohydrate quantity and quality in relation to body mass index. Journal of the American Dietetic Association, 107(10), 1768-1780.

[54] Haimoto, H., Iwata, M., Wakai, K., & Umegaki, H. (2008). Long-term effects of a diet loosely restricting carbohydrates on HbA1c levels, BMI and tapering of sulfonylureas in type 2 diabetes: a 2-year follow-up study. Diabetes Research and Clinical Practice, 79(2), 350-356.

[55] Harding, A. H., Day, N. E., Khaw, K. T., Bingham, S., Luben, R., Welsh, A., & Wareham, N. J. (2004). Dietary fat and the risk of clinical type 2 diabetes: the

European prospective investigation of Cancer-Norfolk study. American Journal of Epidemiology, 159(1), 73-82.

[56] Hodge, A. M., English, D. R., ODea, K., & Giles, G. G. (2004). Glycemic index and dietary fiber and the risk of type 2 diabetes. Diabetes Care, 27(11), 2701-2706.

[57] Hu, G., Lindström, J., Valle, T. T., Eriksson, J. G., Jousilahti, P., Silventoinen, K., ... & Tuomilehto, J. (2004). Physical activity, body mass index, and risk of type 2 diabetes in patients with normal or impaired glucose regulation. Archives of Internal Medicine, 164(8), 892-896.

[58] Kant, A. K., Graubard, B. I., & Atchison, E. A. (2009). Intakes of plain water, moisture in foods and beverages, and total water in the adult US populationnutritional, meal pattern, and body weight correlates: National Health and Nutrition Examination Surveys 19992006. The American Journal of Clinical Nutrition, ajcn-27749.

[59] Kirk, J. K., Graves, D. E., Craven, T. E., Lipkin, E. W., Austin, M., & Margolis, K. L. (2008). Restricted-carbohydrate diets in patients with type 2 diabetes: a meta-analysis. Journal of the American Dietetic Association, 108(1), 91-100.

[60] Lopez-Ridaura, R., Willett, W. C., Rimm, E. B., Liu, S., Stampfer, M. J., Manson, J. E., & Hu, F. B. (2004). Magnesium intake and risk of type 2 diabetes in men and women. Diabetes Care, 27(1), 134-140.

[61] Ma, Y., Olendzki, B., Chiriboga, D., Hebert, J. R., Li, Y., Li, W., ... & Ockene, I. S. (2005). Association between dietary carbohydrates and body weight. American Journal of Epidemiology, 161(4), 359-367.

[62] Malik, V. S., Li, Y., Tobias, D. K., Pan, A., & Hu, F. B. (2016). Dietary protein intake and risk of type 2 diabetes in US men and women. American Journal of Epidemiology, 183(8), 715-728.

[63] Menke, A., Casagrande, S., Geiss, L., & Cowie, C. C. (2015). Prevalence of and trends in diabetes among adults in the United States, 1988-2012. JAMA, 314(10), 1021-1029.

[64] Merchant, A. T., Vatanparast, H., Barlas, S., Dehghan, M., Shah, S. M. A., De Koning, L., & Steck, S. E. (2009). Carbohydrate intake and overweight and obesity among healthy adults. Journal of the American Dietetic Association, 109(7), 1165-1172.

[65] Meyer, K. A., Kushi, L. H., Jacobs, D. R., Slavin, J., Sellers, T. A., & Folsom, A. R. (2000). Carbohydrates, dietary fiber, and incident type 2 diabetes in older women. The American Journal of Clinical Nutrition, 71(4), 921-930.

[66] Nelson, K. M., Reiber, G., & Boyko, E. J. (2002). Diet and exercise among adults with type 2 diabetes. Diabetes Care, 25(10), 1722-1728.

[67] Nguyen, N. T., Nguyen, X. M. T., Lane, J., & Wang, P. (2011). Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 19992006. Obesity Surgery, 21(3), 351-355.

[68] Pan, A., Malik, V. S., Schulze, M. B., Manson, J. E., Willett, W. C., & Hu, F. B. (2012). Plain-water intake and risk of type 2 diabetes in young and middle-aged women. The American Journal of Clinical Nutrition, 95(6), 1454-1460.

[69] Papanikolaou, Y., Brooks, J., Reider, C., & Fugoni 3rd, V. L. (2014). Dietary magnesium usual intake is associated with favorable diabetes-related physiological outcomes and reduced risk of metabolic syndrome: an NHANES 20012010 analysis. Journal of Human Nutrition & Food Science, 2(3), 1038.

[70] Pham, P. C. T., Pham, P. M. T., Pham, S. V., Miller, J. M., & Pham, P. T. T. (2007). Hypomagnesemia in patients with type 2 diabetes. Clinical Journal of the American Society of Nephrology, 2(2), 366-373.

[71] Resnick, L. M., Altura, B. T., Gupta, R. K., Laragh, J. H., Alderman, M. H., & Altura, B. M. (1993). Intracellular and extracellular magnesium depletion in type 2 (non-insulin-dependent) diabetes mellitus. Diabetologia, 36(8), 767-770.

[72] Roussel, R., Fezeu, L., Bouby, N., Balkau, B., Lantieri, O., Alhenc-Gelas, F., Marre, M., Bankir, L. & DESIR Study Group. (2011). Low water intake and risk for new-onset hyperglycemia. Diabetes Care, 34(12), 2551-2554.

[73] Salmeron, J., Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G. A., Rimm, E. B., & Willett, W. C. (2001). Dietary fat intake and risk of type 2 diabetes in women. The American Journal of Clinical Nutrition, 73(6), 1019-1026.

[74] Shang, X., Scott, D., Hodge, A. M., English, D. R., Giles, G. G., Ebeling, P. R., & Sanders, K. M. (2016). Dietary protein intake and risk of type 2 diabetes: results from the Melbourne Collaborative Cohort Study and a meta-analysis of prospective studies. The American Journal of Clinical Nutrition, ajcn140954.

[75] Sluijs, I., Beulens, J. W., Spijkerman, A. M., Grobbee, D. E., & Van der Schouw, Y. T. (2010). Dietary intake of total, animal, and vegetable protein and risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-NL study. Diabetes Care, 33(1), 43-48.

[76] Streiner, D. L. (2005). Finding our way: an introduction to path analysis. The Canadian Journal of Psychiatry, 50(2), 115-122.

[77] Trichopoulou, A., Gnardellis, C., Benetou, V., & Lagiou, P. (2002). Lipid, protein and carbohydrate intake in relation to body mass index. European Journal of Clinical Nutrition, 56(1), 37.

[78] Tsilas, C. S., de Souza, R. J., Mejia, S. B., Mirrahimi, A., Cozma, A. I., Jayalath, V. H., Ha, V., Tawfik, R., Di Buono, M., Jenkins, A.L. & Leiter, L. A. (2017). Relation of total sugars, fructose and sucrose with incident type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies. Canadian Medical Association Journal, 189(20), E711-E720.

[79] Van Dam, R. M., Willett, W. C., Rimm, E. B., Stampfer, M. J.,& Hu, F. B. (2002). Dietary fat and meat intake in relation to risk of type 2 diabetes in men. Diabetes Care, 25(3), 417-424.

[80] Van Dam, R. M., Hu, F. B., Rosenberg, L., Krishnan, S., & Palmer, J. R. (2006). Dietary calcium and magnesium, major food sources, and risk of type 2 diabetes in US black women. Diabetes Care, 29(10), 2238-2243.

[81] Van Nielen, M., Feskens, E. J., Mensink, M., Sluijs, I., Molina, E., Amiano, P., Ardanaz, E., Balkau, B., Beulens, J.W., Boeing, H. & Clavel-Chapelon, F. (2014). Dietary protein intake and incidence of type 2 diabetes in Europe: the EPIC-InterAct Case-Cohort Study. Diabetes Care, 37(7), 1854-1862. Chicago

[82] Wang, Y., Bolge, S. C., Lopez, J. M., Zhu, V. J., & Stang, P. E. (2016). Changes in Body Weight Among People With Type 2 Diabetes Mellitus in the United States, NHANES 2005-2012. The Diabetes Educator, 42(3), 336-345.

[83] World Health Organization (2016). Global report on diabetes. World Health Organization.

[84] Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20(7), 557-585.

[85] Wright, S. (1934). The method of path coefficients. The Annals of Mathematical Statistics, 5(3), 161-215.

[86] Zimmet, P., Alberti, K. G., Magliano, D. J., & Bennett, P. H. (2016). Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. Nature Reviews Endocrinology, 12(10), 616-622.

APPENDIX

Appendix 1

This section includes the additional figures and tables in Chapter 2.

Table A1.1: All possible candidate profiles with nodal parameters for $T = 4$. Nodal parameters for ORI method are defined as the parameters linked to all the other parameters where the two parameters are considered as linked if the inequality between them is pre-specified.

Candidate Profiles	**Nodal Parameters**
$C_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$	-
$C_1 : \mu_1 < \mu_2 < \mu_3 < \mu_4$	$\mu_1, \mu_2, \mu_3, \mu_4$
$C_2 : \mu_1 < \mu_2 < \mu_3 > \mu_4$	μ_3
$C_3 : \mu_1 < \mu_2 > \mu_3 < \mu_4$	None
$C_4 : \mu_1 < \mu_2 > \mu_3 > \mu_4$	μ_2
$C_5 : \mu_1 > \mu_2 < \mu_3 < \mu_4$	μ_2
$C_6 : \mu_1 > \mu_2 < \mu_3 > \mu_4$	None
$C_7 : \mu_1 > \mu_2 > \mu_3 < \mu_4$	μ_3
$C_8 : \mu_1 > \mu_2 > \mu_3 > \mu_4$	$\mu_1, \mu_2, \mu_3, \mu_4$

Table A1.2: Detailed Results of ORI method for ovarian tumor study. P-values for test against null are reported in the last column. * indicates the pTyr site is classified to the same profile using both ORI method and the MAP estimator of Bayesian model.

Protein	pTyr	Sequence	p-values
$C_0 = \{\mu \in R^4 : \mu_1 = \mu_2 = \mu_3 = \mu_4\}$			
annexin A2 isoform 2	ANXA2*	LSLEGDHSTPPSA[y]GSVK	0.1031
glucocorticoid receptor DNA binding factor 1	ARHGAP35*	NEEENI[y]SVPHDSTQGK	0.2196
cell division cycle 2 protein isoform 1	CDK1*	IGEGT[y]GVVYK	0.2036
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2 isoform 1	DYRK2*	VYT[y]IQSR	0.0627
ephrin receptor EphB3 precursor	EPHB3*	FLEDDPSDPT[y]TSSLGGK	0.0813
filamin A, alpha isoform 1	FLNA*	VHSPSGALEEcYVTEIDQDK[y]AVR	0.4594
high density lipoprotein binding protein	HDLBP*	MD[y]VEINIDHK	0.1260
homeodomain-interacting protein kinase 1 isoform 2	HIPK1*	AVcST[y]LQSR	0.0876
phosphoinositide-3-kinase, regulatory subunit 2 (beta)	PIK3R2*	SREYDQL[y]EEYTR	0.2029
serine/threonine-protein kinase PRP4K	PRPF4B*	LcDFGSASHVADNDITP[y]LVSR	0.1589
splicing factor, arginine/serine-rich 9	SRSF9*	G[s]PHYFSPFRPY	0.0526
tensin like C1 domain containing phosphatase isoform 1	TENC1*	GPLDGSP[y]AQVQRPPR	0.3912
tensin	TNS1*	DDGMEEVVGHTQGPLDGSL[y]AK	0.2438
TYRO protein tyrosine kinase binding protein isoform 1 precursor	TYROBP	ITETESP[y]QELQGQR	0.0538
vinculin isoform VCL	VCL*	SFLDSG[y]R	0.2703
$C_1 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 < \mu_4\}$			
myelin protein zero-like 1 isoform a	MPZL1*	SESVV[y]ADIR	0.0073
$C_2 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 > \mu_4\}$			
docking protein 1	DOK1	VKEEGYELPYNPATDD[y]AVPPPR	0.0210
GRB2-associated binding protein 1 isoform a	GAB1*	SSGSGSSVADERVD[y]VVVDQQK	0.0002
SH2 containing inositol phosphatase isoform b	INPP5D	EKL[y]DFVK	0.0048
signal transducer and activator of transcription 3 isoform 1	STAT3	YcRPESQEHPEADPGSAAP[y]LK	0.0142
Wiskott-Aldrich syndrome gene-like protein	WASL*	VI[y]DFIEK	0.0010
$C_3 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 < \mu_4\}$			
mitogen-activated protein kinase 14 isoform 1	MAPK14*	HTDDEMTG[y]VATR	0.0109
$C_4 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 > \mu_4\}$			
None			
$C_5 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 < \mu_4\}$			
None			
$C_6 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 > \mu_4\}$			
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A isoform 3	DYRK1A	IYQ[y]IQSR	0.0238
mitogen-activated protein kinase 7 isoform 1	MAPK7	GLcTSPAEHQYFMTE[y]VATR	0.0054
NKF3 kinase family member	PEAK1	ASTDVAGQAVTINLVPTEEQAKP[y]R	0.0058
$C_7 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 < \mu_4\}$			
NCK adaptor protein 1	NCK1*	L[y]DLNMPAYVK	0.0028
$C_8 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 > \mu_4\}$			
cyclin-dependent kinase-like 5	CDKL5*	NLSEGNNANYTE[y]VATR	0.0019
ephrin receptor EphA2	EPHA2*	VLEDDPEAT[y]TTSGGKIPIR	0.0004
FYN binding protein (FYB-120/130) isoform 2	FYB*	TTAVEID[y]DSLK	0.0217
neural precursor cell expressed, developmentally down-regulated 9 isoform 1	NEDD9	EKD[y]DFPPPMR	0.0020
protein kinase C, delta	PRKCD*	RSDSASSEPVGI[y]QGFEK	0.0071
paxillin	PXN	VGEEEHV[y]SFPNK	0.0010

Table A1.3: Selected profiles and the percentage of $J = 35$ sites assigned to each profile for $T = 5$ and $T = 8$ in simulation study of Section 2.4.2.

Selected Profiles	% Assigned to Profile
$T = 5$	
$C_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	20%
$C_1 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5$	10%
$C_2 : \mu_1 < \mu_2 > \mu_3 > \mu_4 > \mu_5$	10%
$C_3 : \mu_1 < \mu_2 < \mu_3 > \mu_4 > \mu_5$	5%
$C_4 : \mu_1 < \mu_2 < \mu_3 < \mu_4 > \mu_5$	10%
$C_5 : \mu_1 > \mu_2 < \mu_3 < \mu_4 < \mu_5$	5%
$C_6 : \mu_1 > \mu_2 > \mu_3 < \mu_4 < \mu_5$	5%
$C_7 : \mu_1 > \mu_2 > \mu_3 > \mu_4 < \mu_5$	10%
$C_8 : \mu_1 > \mu_2 > \mu_3 > \mu_4 > \mu_5$	10%
$C_9 : \mu_1 < \mu_2 < \mu_3 > \mu_4 < \mu_5$	5%
$C_{10} : \mu_1 < \mu_2 > \mu_3 < \mu_4 < \mu_5$	10%
$T = 8$	
$C_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8$	20%
$C_1 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6 < \mu_7 < \mu_8$	5%
$C_2 : \mu_1 < \mu_2 > \mu_3 < \mu_4 < \mu_5 < \mu_6 < \mu_7 < \mu_8$	5%
$C_3 : \mu_1 < \mu_2 > \mu_3 < \mu_4 > \mu_5 < \mu_6 < \mu_7 < \mu_8$	5%
$C_4 : \mu_1 < \mu_2 < \mu_3 < \mu_4 > \mu_5 > \mu_6 > \mu_7 > \mu_8$	10%
$C_5 : \mu_1 < \mu_2 < \mu_3 < \mu_4 > \mu_5 < \mu_6 > \mu_7 > \mu_8$	5%
$C_6 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6 > \mu_7 > \mu_8$	5%
$C_7 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6 < \mu_7 > \mu_8$	10%
$C_8 : \mu_1 > \mu_2 > \mu_3 < \mu_4 < \mu_5 > \mu_6 < \mu_7 < \mu_8$	5%
$C_9 : \mu_1 > \mu_2 > \mu_3 < \mu_4 < \mu_5 < \mu_6 > \mu_7 < \mu_8$	5%
$C_{10} : \mu_1 > \mu_2 > \mu_3 > \mu_4 < \mu_5 < \mu_6 < \mu_7 < \mu_8$	5%
$C_{11} : \mu_1 > \mu_2 > \mu_3 > \mu_4 > \mu_5 < \mu_6 > \mu_7 < \mu_8$	5%
$C_{12} : \mu_1 < \mu_2 > \mu_3 > \mu_4 > \mu_5 > \mu_6 < \mu_7 > \mu_8$	5%
$C_{13} : \mu_1 > \mu_2 > \mu_3 > \mu_4 < \mu_5 > \mu_6 < \mu_7 < \mu_8$	5%
$C_{14} : \mu_1 > \mu_2 > \mu_3 > \mu_4 > \mu_5 > \mu_6 > \mu_7 > \mu_8$	5%

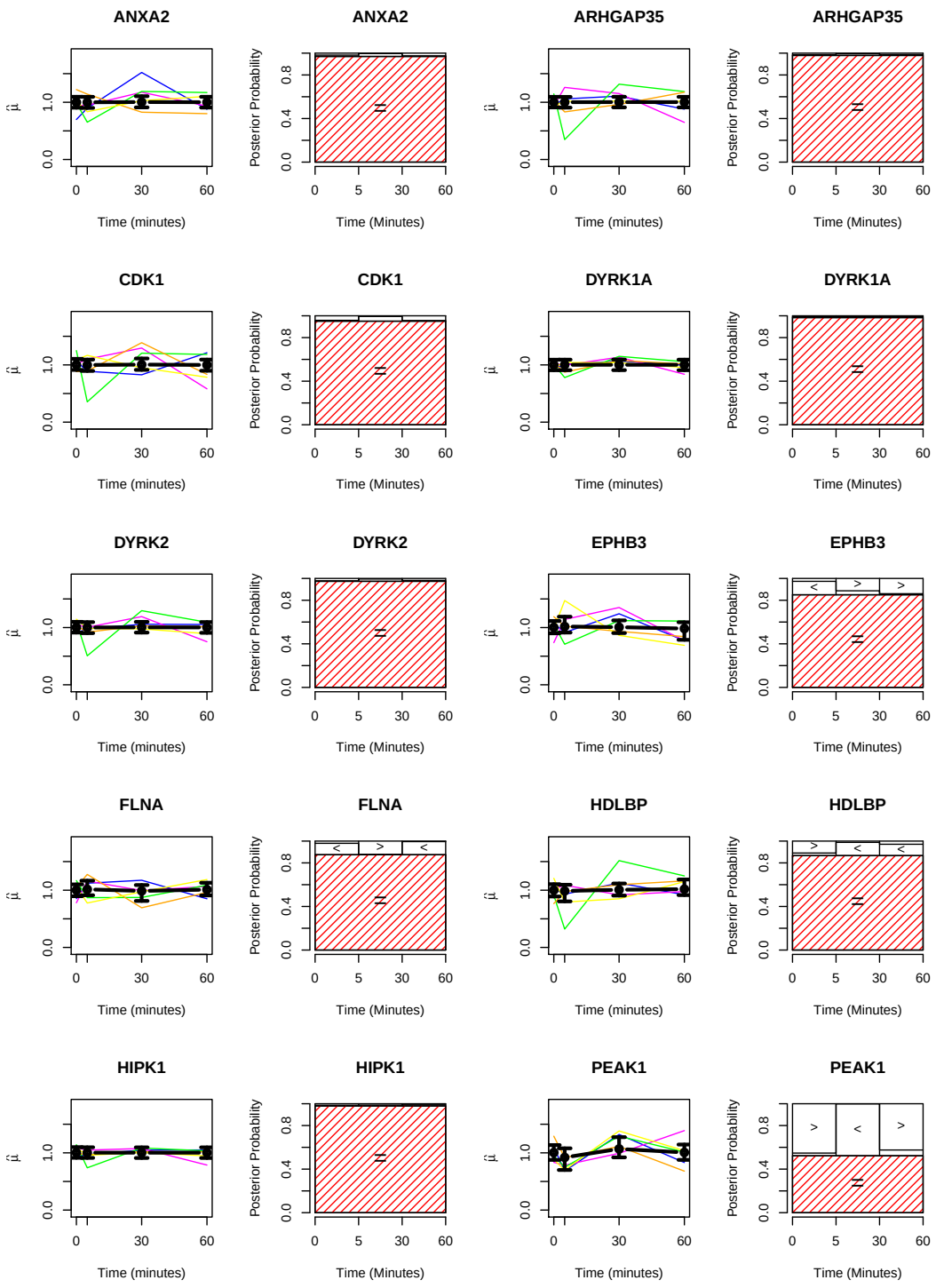

Figure A1.1: Estimated Trajectory with 95% credible interval (1*st* and 3*rd* columns) and MPW estimators (2*nd* and 4*th* columns) of proteins 1-10. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

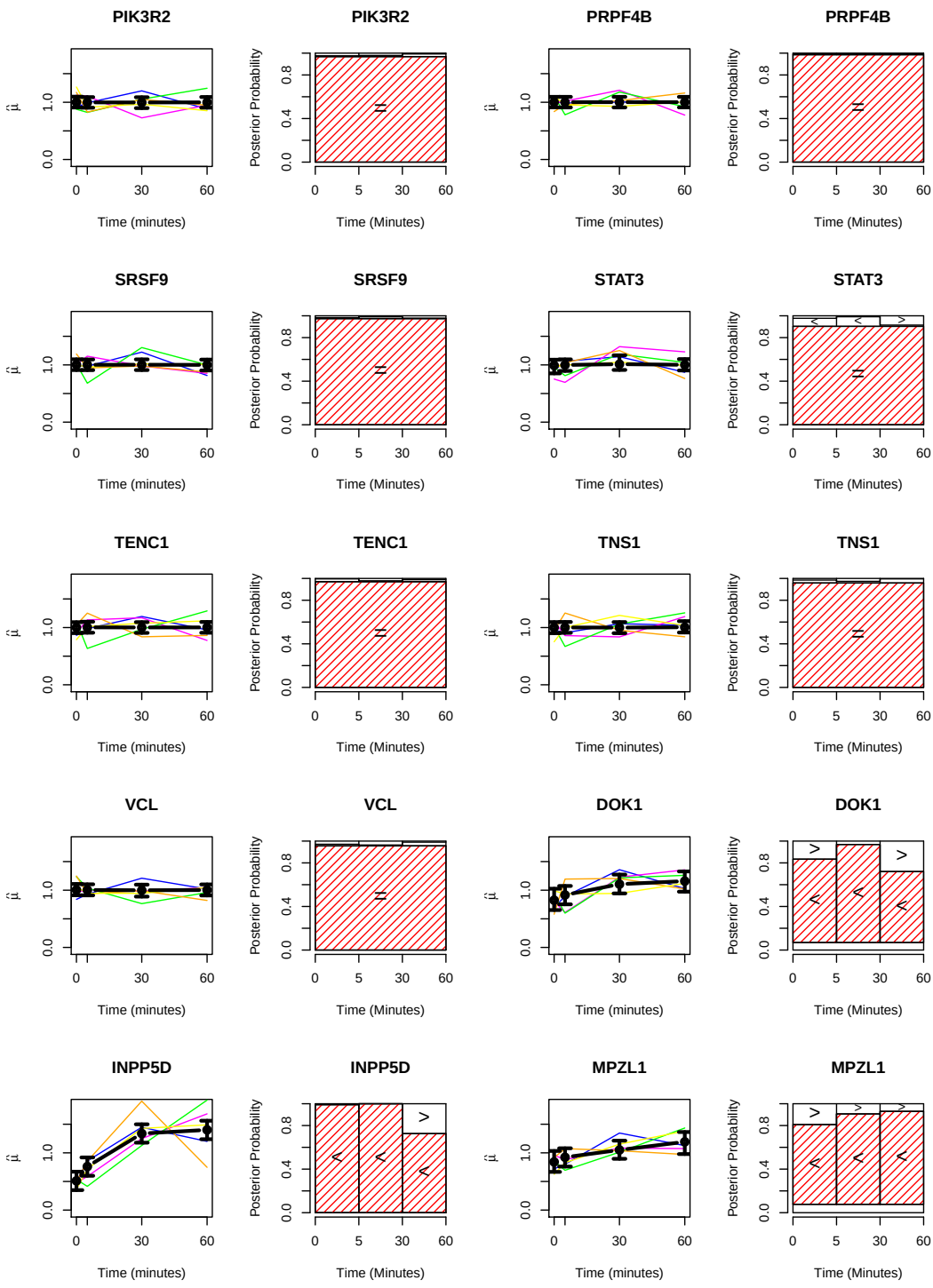

Figure A1.2: Estimated Trajectory with 95% credible interval (1*st* and 3*rd* columns) and MPW estimators (2*nd* and 4*th* columns) of proteins 11-20. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

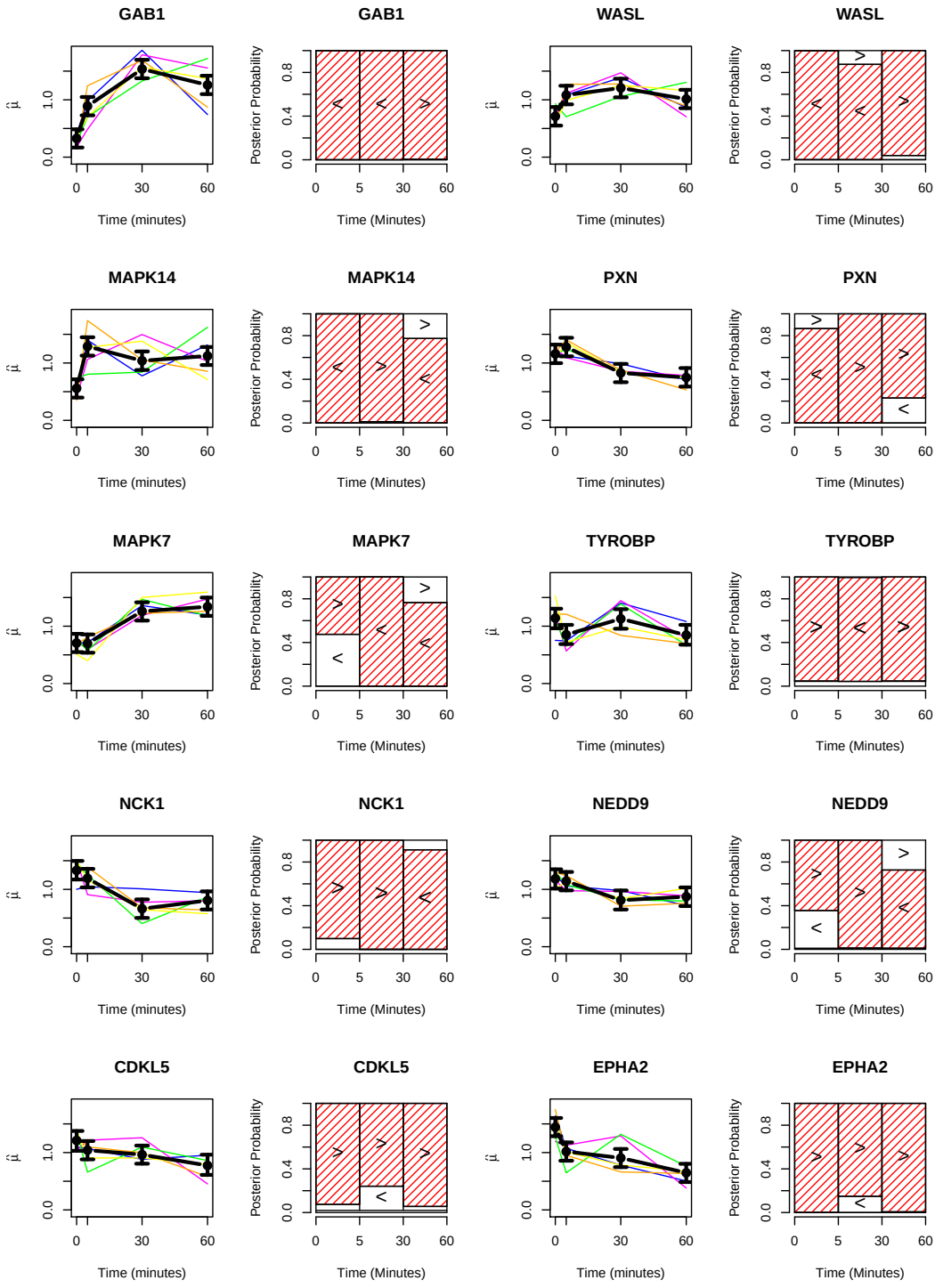

Figure A1.3: Estimated Trajectory with 95% credible interval (1*st* and 3*rd* columns) and MPW estimators (2*nd* and 4*th* columns) of proteins 21-30. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

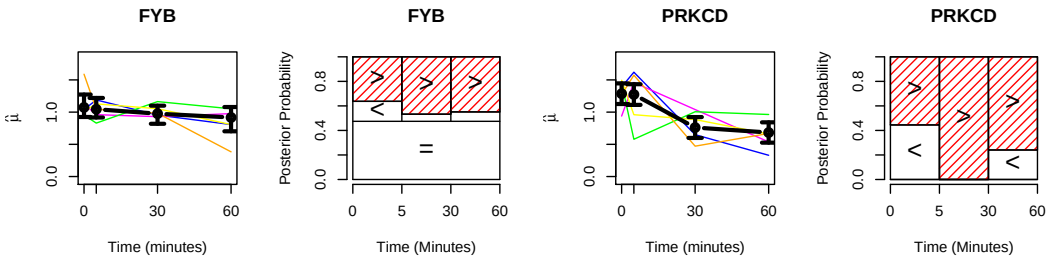

Figure A1.4: Estimated Trajectory with 95% credible interval ($1st$ and $3rd$ columns) and MPW estimators ($2nd$ and $4th$ columns) of proteins 31-32. The colored lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

Appendix 2

This section includes the derivation of the estimated variance of the superiority score obtained from the propensity score-based matching in Chapter 3.

Let us denote $I_{1i} = I(Y_i^{(1)} > Y_i^{(0)})$ and $I_{0i} = I(Y_i^{(1)} = Y_i^{(0)})$. Then in the propensity score-based matching method, the estimated superiority score can be expressed as:

$$\begin{aligned}\hat{\gamma}_{Mat} &= \frac{1}{N}\sum_{i=1}^{N}[I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2}I(Y_i^{(1)} = Y_i^{(0)})] \\ &= \frac{1}{N}\sum_{i=1}^{N}[I_{1i} + \frac{1}{2}I_{0i}]\end{aligned}$$

Then the estimated variance for $\hat{\gamma}_{Mat}$ is derived as follows.

$$\begin{aligned}\hat{\sigma}_{\gamma,Mat} &= Var(\hat{\gamma}_{Mat}) \\ &= Var(\frac{1}{N}\sum_{i=1}^{N}[I_{1i} + \frac{1}{2}I_{0i}]) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}Var(I_{1i} + \frac{1}{2}I_{0i}) \\ &= \frac{1}{N}Var(I_{11} + \frac{1}{2}I_{01}) \\ &= \frac{1}{N}\{E(I_{11} + \frac{1}{2}I_{01})^2 - [E(I_{11} + \frac{1}{2}I_{01})]^2\} \\ &= \frac{1}{N}\{E(I_{11}) + \frac{1}{4}E(I_{01}) - [E(I_{11} + \frac{1}{2}I_{01})]^2\} \\ &= \frac{1}{N}\{\frac{1}{N}\sum_{i=1}^{N}I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{4}I(Y_i^{(1)} = Y_i^{(0)}) \\ &- [\frac{1}{N}\sum_{i=1}^{N}(I(Y_i^{(1)} > Y_i^{(0)}) + \frac{1}{2}I(Y_i^{(1)} = Y_i^{(0)}))]^2\} \\ &= \frac{1}{N}(\hat{p}_1 + \frac{1}{4}\hat{p}_0 - \hat{\gamma}_{Mat}^2).\end{aligned}$$

where $\hat{p}_1 = \frac{1}{N}\sum_{i=1}^{N}I(Y_i^{(1)} > Y_i^{(0)})$ and $\hat{p}_0 = \frac{1}{N}\sum_{i=1}^{N}I(Y_i^{(1)} = Y_i^{(0)})$

CURRICULUM VITA

NAME: You Wu

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40292

EDUCATION: Bachelor of Sceince,
Industrial and Commercial College of Hebei University, 2011
Masters in Applied Statistics,
University of Peradeniya, 2013

PUBLICATIONS: Wu, Y., Gaskins, J., Kong, M. and Datta, S.(2017).
Profiling the Effects of Short Time-course Cold Ischemia on
Tumor Protein Phosphorylation using a Bayesian Approach.
Biometrics. (Accepted)

Wu, Y., Datta, S., Little, B. and Kong, M.
Study of different statistical methods for estimating treatment
effects when outcome variable is ordinal and confounding exists
(In process)

Wu, Y., Little, B., Datta, S. and Kong, M.

Causal Analysis of Dietary Information and Physical Activity in Type 2 Diabetes by Gender in White, African American and Mexican American: National Health and Nutrition Examination Surveys 2011-2014
(In process)

PRESENTATIONS: ASA-KY Chapter Meeting, Lexington, KY, April 2016.
Profiling the effects of short time-course cold ischemia on tumor protein phosphorylation using a Bayesian approach.

The Joint Statistical Meeting (JSM), Chicago, IL,
July 30- August 4, 2016. Contributed poster presentations:
Profiling the effects of short time-course cold ischemia on tumor protein phosphorylation using a Bayesian approach.

Seminar Series, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA, April 14, 2017.
Study of different statistical methods for estimating treatment effects when outcome variable is ordinal and confounding exists.

ASA-KY Chapter Meeting, Louisville, KY, April 2017.
Profiling the effects of short time-course cold ischemia on tumor protein phosphorylation using a Bayesian approach.

The Southern Regional Council on Statistics (SRCOS)
Summer Research Conference (SRC), Jekyll Island, GA,
July 4-July 7, 2017.

NSF/ Harshbarger Student Poster Session:
Study of different statistical methods for estimating treatment effects when outcome variable is ordinal and confounding exists.

HONORS AND AWARDS

University Graduate Scholarship, 2011-2013,
University of Cincinnati

Travel Award, 2016, Graduate school,
University of Louisville

NSF funded Harshbarger Travel award for the SRCOS Summer Research Conference, Jekyll Island, GA, 2017