STATISTICAL INFERENCE WITH MULTIPLE DATA SOURCES

by

Parichoy Pal Choudhury

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

January, 2016

Abstract

In this dissertation, I develop statistical methods to address three important scientific problems. A common theme behind these methods is the stitching together of multiple data sources to address the scientific questions of interest. In the first paper (Chapter 2), I propose a novel statistical framework to learn about the association between a secondary outcome (e.g., obesity) and a genetic risk factor (e.g., ORMDL3 locus on Chromosome 17) from a genetic case-control study based on asthma. The method involves the use of asthma prevalence information from a relevant sample survey. In the second paper (Chapter 3), I develop a method to evaluate whether there are adverse health consequences of kidney donation. To address this question, I use data on donors from the Wellness and Health Outcomes in LivE Donors (WHOLE-DONOR) Study and on healthy non-donors from the Atherosclerosis Risk in Communities (ARIC) and Coronary Artery Risk Development in Young Adults (CARDIA) studies. In the third paper (Chapter 4), I propose a covariate-adjusted method for testing the difference between two treatment groups where the measured outcome is a function. The proposed method utilizes information from repeated mea-

ABSTRACT

sures of daily oxygen consumption function and scalar body composition measures (e.g., lean mass, fat mass) on two groups of mice, one with and one without a specific gene.

Primary Readers: Daniel O. Scharfstein, Ciprian M. Crainiceanu, Allan Massie and Eliseo Guallar

Secondary Readers: Michael Rosenblum and Ramin Mojtabai.

Acknowledgments

I take this opportunity to express my sincere gratitude to the people who have helped me arrive at this significant moment of my life. First of all, I would like to thank my research advisor, Dr. Daniel O. Scharfstein, for the incredible mentorship that I have received from him during my PhD study. Under his guidance, my eyes were opened to the vast potential of statistics and the role played by statisticians in the scientific discovery process. It was my privilege to work with such an outstanding statistical scientist, considered a leader in his field. As a result, I learned to be rigorous, skeptical and perform research with integrity. I value the friendship and rapport I have with him above and beyond academics: when times were particularly challenging his usage of remarkable phrases like “hang in there” and “take it easy” was immensely helpful to boost my morale; something I have used with success when providing guidance to friends in difficult situations. I cherish the fond memories of intense academic discussion in meetings and light hearted humor during walks to the train station.

I am grateful to Drs. Ciprian M. Crainiceanu, Allan Massie and Eliseo Guallar for serving on my thesis committee. In particular, I would like to thank Dr. Ciprian M. Crainiceanu for providing me guidance with my third thesis paper. As a result, I was exposed to the exciting area of research in functional data analysis. In this connection, a special topics course on Applied Functional Data Analysis with Drs. Luo Xiao and Vadim Zippunikov and informal discussions with Dr. Luo Xiao were particularly helpful. A fantastic collaboration with the research group of Drs. Allan Massie and Dorry Segev resulted in my second thesis paper. The group meetings and informal discussions with Dr. Allan Massie helped me understand and develop clarity on the scientific background pertaining to the problem.

I remain grateful to our Department Chair, Dr. Karen Bandeen-Roche for providing invaluable advice whenever needed. Be it an academic problem or an administrative concern, she was always available to help find solutions. Through her connections, I had a productive research experience with the Center of Aging and Health at JHU where I had a chance to work with some wonderful statisticians and scientists: Drs. Qian-li Xue, Paulo Chaves, Ravi Varadhan, Reyhan Westbrook and Rita Kalyani. These collaborations were instrumental in developing my skills on the more applied side of statistics.

Over the last few years I had the privilege of working with some outstanding researchers: Drs. Saunak Sen (University of Tennessee, Memphis), Nilanjan Chatterjee (JHU), Orestis Panagiotou (National Cancer Institute), Brian Schwartz (JHU), Aravinda Chakravarti (JHU), Ivan Diaz (Google), Kevin Psoter (JHU), Yi Lu (JHU) and Chris McMahan (Clemson University). I thank each one of them for helping me develop my scientific and statistical expertise in connection to complex research problems. Following my PhD, I am excited to join Dr. Nilanjan Chatterjee's research group as a postdoctoral fellow to work at the interface of causal inference and statistical genetics. Special thanks are due to him for providing me this opportunity.

The departmental workgroups were extremely helpful for facilitating research related discussion and brainstorming. I enjoyed going to the Causal Inference meetings. They helped shape my understanding of this very exciting area of research. I have also learned a lot from the meetings of the SLAM and SMART groups. In recent times, a Statistical Genetics workgroup is developing under the leadership of Dr. Nilanjan Chatterjee and we have had some productive meetings. This is particularly exciting for me as I have a strong interest in learning about the statistical methods and applications in genetics research. Overall, the workgroups provide a vibrant environment for stimulating research discussions and also help socially connect researchers working on similar problems.

The Departmental Staff have been incredibly helpful to me throughout my PhD study. I would like to make special mention of Mary Joy Argo, Debbie Cooper and Jiong Yang. Be it an academic matter (e.g., course requirement, forming examination committees), Mary Joy was always around to share her expertise. She seemed to have an answer within seconds to any academic issue. Debbie Cooper was extremely friendly and helpful as a point person for administrative concerns. She would always greet me with a big smile and patiently explain each and every detail. I have also got a tremendous amount of help from Jiong to resolve some of the computing issues. I express my sincere thanks to the Department of Biostatistics for providing me financial support during my PhD study.

Over the last few years, my experience of living in Baltimore and getting through this challenging PhD program have been truly life changing. The "Charm City", as it is popularly known, helped me stand on my feet and transform into an independent person. I learned how to pick myself up after failing on numerous occasions and had the life realization that any kind of success involves patiently getting through the failures with renewed vigor. I had the opportunity of making some great friends here in Baltimore. Many of them are among the students, postdocs and alumni of our department, who have been incredible in more ways than one. I would like to thank Sumit Sahu, Shatabdi Pal and Prosenjit Kundu for helping me find a home away from my home. I value their friendship and support especially towards the finishing stages of my study. Some other friends who deserve special mention are Anuradha Dutta, Anusha Chatterjee, Debmalya Chakraborty, Emily Huang, Fang Han, Haoyu Zhang, Jiawei Bai, Oindrila Dawn, Rahul Bhattacharya, Sayantan Datta, Sharmistha Mukherjee, Snehashis Roy, Souradyuti Ghosh, Suman Dasgupta, Yenny Webb-Vargas, Zhenke Wu and others. The Indian Graduate Students' Association at JHU, popularly known as JHU-IGSA, is extremely pro-active in organizing social events and meetings involving the vibrant Indian student community on campus. Not

only did these events provide me with a welcome break from the stressful graduate life, but I have also made some outstanding new friends through this organization.

My family: Dr. Pabitra Pal Choudhury (father), Chandrima Pal Choudhury (mother), Prakriti Pal Choudhury (sister), Rekha Paul (maternal grandmom) have been extremely supportive throughout my PhD study. My father is a professor at the world renowned Indian Statistical Institute. It was his dream that both of his children would finish doctoral studies. He took special care to ensure we had a very productive academic environment at home and obtain the best education at the high school and college level. I was fortunate to receive my pre-college education from South Point School, one of the premier schools in Eastern India. The dedicated teachers made sure I received cutting edge education and was prepared for my next dream: to pursue a five year undergraduate and post graduate training in Mathematics and Statistics at the Indian Statistical Institute. It was an awesome experience to be trained by a team of renowned and dedicated professors at the ISI. Both my high school education and training at the ISI ensured that I had enough preparation to take up my next enormously challenging venture: PhD in Biostatistics at JHU. Special thanks are due to the long list of friends that I have made from these two places, many of whom are around the United States: Abhishek Dan, Agniva Som, Apara Banerjee, Arkaprabha Sarangi, Ipshita Bhattacharya, Nilabja Guha, Oishik Sen, Pramita Bagchi, Pratyaydipta Rudra, Sandipan Roy, Sayantan Banerjee, Sebanti Sengupta, Somak Dutta, Souvick Chatterjee, Sudip Paul, Suvojit Ghosh, Swarnava Mukhopadhyay and many

others whom I have probably missed out; life in the last few years would have been lot less fun without them.

As I get ready to graduate, I would like to gratefully acknowledge the contributions of my family, relatives, friends and the education at South Point, ISI and JHU toward this very exciting journey.

Dedication

This thesis is dedicated to my family, friends and relatives.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Multiple data sources are often required to address important scientific questions. A single data source is often insufficient to estimate a target parameter of interest. A classic example is in genetic epidemiology where interest focuses on the causal effect of a biomarker (e.g. Vitamin D) on a disease endpoint (e.g. Multiple Sclerosis) using the genes predictive of the biomarker as instrumental variables. A single dataset does not usually contain information on the outcome, biomarker and genes and even when it does the sample size is usually too low to yield precise inferences. Typically one learns about the instrument-biomarker association from one data source and instrument-gene association from another data source and these data sources are married to estimate the causal effect of interest (Mokry *and others*, 2015; Burgess *and others*, 2015). The environmental epidemiology literature also contain examples where interest focuses on the association between an environmental exposure and health outcomes. Typically, the exposure and health information are obtained from distinct data sources; confounding factors may even be obtained from a third data source. For example, Dominici *and others* (2006) studied the association between fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Health information was obtained from billing claims of Medicare enrollees, pollution data were obtained from EPA's Aerometric Information Retrieval Service and weather information (confounders) were obtained from the National Climatic Data Center on the Earth-Info CD database. Social scientists also use multiple data sources in their research. For example, Corvalan *and others* (2015) discusses how to construct bounds on the causal effect of a change in a Chilean electoral law on voter turnout using two separate data sources: aggregate level data of voter counts and individual level demographic data.

This dissertation is devoted to the development of statistical methods to address three important scientific problems. The methods involve combining information from multiple data sources to address the scientific questions of interest. The following three sections provide a gentle introduction to these problems.

1.1 Enhancing Genetic Case-Control Studies Using Sample Surveys

In a typical case-control study, individuals are ascertained on the basis of their disease status, i.e. whether they are a case or a control. The study design is retrospective in the sense that exposure information is collected retrospectively. This design is useful for characterizing the association of an exposure of interest and the case-control status. In a genetic case-control study, the exposure of interest is usually a genetic variant. The disease status that determines whether an individual is a case or a control is often called a *primary phenotype*. In these studies, it is common of investigators to collect a battery of additional health outcomes referred to as *secondary phenotypes*. An investigator may often be interested in exploring the relationship between a genetic variant of interest and a secondary phenotype. For example, one may want to learn about the relationship between a gene of interest and obesity from an asthma case-control study that also collects obesity information for each individual. Since case-control data are not a random sample from the target population, the observed association between a genetic risk factor and a secondary phenotype may be biased. In order to correct for this bias it is necessary to utilize external information or assumptions. In contrast to the case-control study design, a sample survey provides representative information on the target population of interest. While existing methods make additional assumptions which may not be plausible in a given scientific setting, we propose an inferential framework that combines information from a case-control study and a sample survey from the target population to learn about the association between a secondary phenotype and a genetic risk factor. In particular, the sample survey helps us obtain point estimates and uncertainty of the conditional (on covariates) prevalence of disease that determines the case-control status. We can learn about the conditional distribution of the secondary phenotype and genetic risk factor given the primary phenotype and covariates from the case-control study. Using both data sources, the conditional distribution of the secondary phenotype and the gene given covariates becomes estimable.

By way of illustration, we study the relationship between a candidate gene (i.e., *IKZF3-ZPBP2-GSDMB-ORMDL3* locus on chromosome 17q21) and obesity and how this relationship differs by ethnicity (i.e., Puerto Ricans vs Mexicans). We use data from the GALA (Genes-Environments and Admixture in Latino Americans) II asthma case-control study and the NHIS (National Health Interview Survey). Our results show that a naive analysis using the case-control data alone does not indicate a gene-obesity association, while the combined analysis indicates a significant recessive association. Moreover, there is no statistically significant evidence in favor of a differential association across ethnicities.

1.2 Causal Effect Among The Exposed: Multiple Data Sources and Censored Outcomes

We develop an inferential framework for estimating the causal effect among "exposed" subjects on a time-to-event outcome, based on multiple data sources and censored outcome information. Our major contribution is to conceptualize a hypothetical point exposure study where subjects are enrolled and allowed to select their own exposure. Using information from two data sources (one from exposed subjects and one from non-exposed subjects with multiple examination times), we describe a process of manufacturing a dataset that closely mimics this hypothetical study. The identification of the causal effect relies on a no unmeasured confounding assumption based on covariates available at exposure selection and a non-informative censoring assumption. Estimation proceeds by fitting separate proportional hazards regression models for exposed and non-exposed subjects using the manufactured dataset and using G-computation to estimate, for exposed subjects, the distributions of time-to-event under exposure and non-exposure. Using these estimated distributions, we compute a parsimonious measure of the causal effect of interest.

We illustrate our methodology by addressing the question of whether kidney donors are putting themselves at increased risk of adverse health consequences. We use information available on live kidney donors derived from hospital records at the Johns Hopkins Hospital and follow-up interviews and healthy non-donors from two prospective cohort studies (i.e., Atherosclerosis Risk in Communities (ARIC) and Coronary Artery Risk Development in Young Adults (CARDIA) studies). We consider two separate endpoints: hypertension-free survival and diabetes-free survival. Our analysis does not provide any significant evidence that kidney donors are putting themselves at an increased risk for these diseases. We also perform a realistic simulation study to evaluate the performance of our proposed methodology.

1.3 Testing Equality of Curves After Covariate Adjustment

We develop simple methodological approaches for global and local tests of the difference between the mean of treatment and control groups when the measured outcome is a function. Our approach utilizes information from two data sources: one coming from subjects in the treatment group and the other coming from the subjects in the control group. The added complexity is that for every subject we have repeated samples for the same curve and additional covariates of interest. A key feature of our proposed methodology is that we are working with covariate adjusted curves which is of critical importance in many applications where the distribution of the covariates differ between groups. We propose a permutation based approach to test for equality of the averages of two functional processes after covariate adjustment. The within group averages are estimated by modeling the relationship of the functional outcome on the covariate using functional regression methods and then averaging with respect to the covariate distribution in each group. The test statistic is the L^2 area under the squared difference curve. We also test for the localized differences between the two average curves using a nonparametric bootstrap of subjects to obtain the 95% pointwise and joint confidence intervals for the difference (Crainiceanu *and others*, 2012).

We illustrate our method by studying the differences in time varying oxygen consumption between Interleukin 10tm1Cgn (IL10tm) mice and wildtype mice after adjusting for body composition measures. While the body weight normalized oxygen consumption is significantly altered in the 10tm1Cgn (IL10tm) mice compared to the wildtype mice; the difference is not significant after adjusting for the ratio of fat and lean mass measured at baseline. This is true for both the global differences and the localized differences. Extensive simulation studies illustrate that the proposed tests preserve the type one errors and are highly sensitive to detecting departures from the null assumption.

1.4 Overview of Dissertation

The dissertation is organized as follows. Chapters 2-4 discuss the details of the projects described in Sections 1.1 - 1.3 above. In each chapter, we present (i) an overview of the scientific problem, the shortcomings of the existing statistical methodology to address that problem and the relevance of our contribution; (ii) a description of the theoretical details of the development of our statistical methods; (iii) an explanation of the scientific findings; and (iv) a general discussion on the scope and limitations of the method and its applicability to problems of similar nature. Chapter 5 is devoted to concluding remarks.

Chapter 2

Enhancing Genetic Case-Control Studies Using Sample Surveys

2.1 Introduction

Consider an unmatched case-control study in which diseased (cases) and non-diseased (controls) individuals are each randomly sampled from a target population. The disease of interest is considered the *primary phenotype*. Further, suppose the main purpose of the study is to discover whether there is an association between genetic factors and the primary phenotype. Towards this end, genetic information is collected on each participant. Investigators often collect a battery of additional clinically relevant phenotypes, referred to as *secondary phenotypes*. Such data are often used to study genetic associations with secondary phenotypes.

CHAPTER 2. ENHANCING GENETIC CASE-CONTROL STUDIES USING SAMPLE SURVEYS

While disease-genotype associations (on an odds ratio scale) can be estimated from case-control data (Prentice and Pyke, 1979), estimating the association between a secondary phenotype and a genotype can only be done with additional information or assumptions. In fact, nominal measures of association between the secondary phenotype and genotypes can be biased (Lee *and others*, 1997). One may observe an association between the secondary phenotype and the genotype in the case-control sample, even if none exists in the population. On the other hand, one may observe no association in the case-control sample, even if one exists in the population.

This issue has received attention in the literature, and several solutions have been proposed. Nagelkerke *and others* (1995) and Kraft (2007) assume that either (a) disease status is conditionally independent of the genotype given the secondary phenotype or (b) disease status is conditionally independent of the secondary phenotype given the genotype. Unfortunately, these assumptions may not hold in the population of interest. A number of methods are available when the sampling fractions for cases and controls are known (Lee *and others*, 1997; Reilly *and others*, 2005; Jiang *and others*, 2006; Richardson *and others*, 2007; Monsees *and others*, 2009; Tchetgen, 2014). Wang and Shete (2011), Chen *and others* (2013), He *and others* (2011), Ghosh *and others* (2013) and Wei *and others* (2013) assume the disease prevalence is known. One needs to be careful when assuming the prevalence is known. Specifically, the prevalence needs to be computed from a population where the conditional distribution of key risk factors given primary phenotype matches that in the case-control study. These methods also neglect the uncertainty in knowledge of prevalence. Li *and others* (2010) and Wei *and others* (2013) considered the case where the disease is rare. Unless the disease is rare, the likelihood-based approach of Lin and Zeng (2009) may be unstable without additional information about prevalence. In short, unless the disease is rare, current methods for analyzing secondary phenotype associations use assumptions that may be false (e.g., conditional independence) or known imprecisely (e.g., prevalence).

We address the problem by obtaining external information from a sample survey of the target population of interest that also measures the primary phenotype. Our inferential framework uses the point estimate and uncertainty of the disease prevalence conditional on covariates from the sample survey. To illustrate our approach, we study the relationship between a candidate gene (associated with asthma) and obesity, and how this relationship differs by ethnicity. We use data from the Genes-Environments & Admixture in Latino Asthmatics (GALA) II study, an asthma case-control study in Latino American children, and the National Health Interview Survey (NHIS) 2010, a national sample survey of households. The GALA II study provides information about the conditional distribution of the genotype, obesity, and key confounders given asthma status and ethnicity; the NHIS study provides information about the probability of asthma given ethnicity and the key confounders. Information from these two distinct data sources are combined to estimate standardized associations between the gene and obesity within ethnicity strata; these are then compared across ethnicities.

2.2 Motivating Example and Data Sources

The GALA II study is the largest pediatric asthma genetic study in US Latinos. The study enrolled approximately equal numbers of cases (children aged 8-21 years with asthma) and controls from five urban cities in the US and Puerto Rico. The two predominant ethnicities of US Latinos are Mexican (63%) and Puerto Rican (9%), who have very different rates of asthma. Approximately 30% of Puerto Rican and 12% of Mexican youth suffer from asthma (`http://www.cdc.gov/asthma/nhis/2011/table2-1.htm`). These prevalences represent two extremes among major ethnic groups in the US. The causes underlying this disparity have puzzled researchers; it is likely that social, cultural, and genetic factors contribute (Thakur *and others*, 2013). Given the disparity in asthma prevalence between Latino subgroups, GALA II was designed to study environmental and genetic factors affecting asthma in Latinos. In addition to genetic data, GALA II collected information on obesity, age, gender and ethnicity.

Another growing public health concern in pediatric populations is obesity whose incidence has been increasing steadily. Many studies have indicated that obesity increases the prevalence and incidence of asthma. Both diseases may arise in childhood, and there are reasons to believe that there are shared etiologic factors that contribute to both diseases (e.g., inflammation). It is also possible that one condition may adversely affect the other (e.g., lung volume is reduced by obesity which leads to reduced lung function). It is, therefore, of interest to examine to what extent common genetic factors contribute to both diseases.

The most prominent genetic region that has been repeatedly implicated in asthma is the *IKZF3-ZPBP2-GSDMB-ORMDL3* locus on chromosome 17q21. The association has been replicated in diverse populations from Europe, North America and Asia. Due to strong linkage disequilibrium across the *17q21* locus, separating the contributions of the genes underlying this locus has been challenging. Nonetheless, because of the co-occurrence of asthma and obesity, this locus is a prime candidate for being associated with obesity susceptibility as well.

We focus on one SNP, *rs12232497*, which has the highest odds ratio for asthma susceptibility in the GALA II population. We examine its association with obesity, being open to the possibility that the association may differ in different ethnic groups (i.e., Mexicans and Puerto Ricans). Our goal is to estimate the association between this SNP and obesity, separately for Mexicans and Puerto Ricans, and evaluate whether there is a differential association.

To realize this goal, we obtain external information from the NHIS-2010 (NCHS, 2011). The NHIS is conducted annually by the National Center for Health Statistics and Centers for Disease Control and Prevention. The NHIS administers face-to-face interviews in a nationally representative sample of households. Within each sampled household with children under the age of 18 years, a detailed survey was conducted on one randomly selected child. A knowledgable adult provided proxy responses for the selected child. Information collected included health measures such as asthma and obesity and demographic factors such as age, gender and ethnicity. Survey weights are included in the NHIS data files to allow for population-level inference.

2.3 Methods

Let A denote asthma status (primary phenotype of interest), for which the case-control sample was assembled. We wish to study the association between genotype (G; coded as 0, 1, 2 based on the number of copies of the minor allele) and the secondary phenotype, obesity (O) within ethnicity strata (E). It is important to control for demographic factors (X), such as age and gender. If the association between G and O is modified by X and the distribution of X is different across E, then we may see a differential association between G and O across strata that results solely from differences in the distribution of X. To address this problem, we seek to estimate the association between the genotype and obesity in a "world" in which the distribution of age and gender is common across strata and is fixed. This is akin to the idea of standardization in epidemiology. We assume the reference population to have uniform age distribution between ages 8-18 years and a 1:1 gender ratio. This is a reasonable approximation to a stable population with low levels of child mortality.

Our goal is the measure the ethnicity-specific association between O and G. In particular, we want the ethnicity-specific joint distribution of O and G, which can be expressed as:

$$P_e[O = o, G = g] = \int P_e[O = o, G = g|X = x]dF(x)$$

where P_e denotes a probability distribution conditional on $E = e$ and $F(x)$ denotes the distribution of demographic factors (age and gender) in the reference population.

Note that, $P_e[O = o, G = g|X = x]$ is not estimable from the case-control data alone or the survey data alone. This is because the survey does not contain genotype information, and the case-control study only allows us to learn about ethnicity-specific joint distributions conditional on asthma and covariates i.e., $P_e[O = o, G = g|A = a, X = x]$ for $a = 0, 1$. However, we can express $P_e[O = o, G = g|X = x]$ as:

$$\begin{aligned} &P_e[O = o, G = g|X = x] \\ &= \sum_{a=0}^{1} P_e[O = o, G = g|A = a, X = x]P_e[A = a|X = x] \\ &= \sum_{a=0}^{1} \underbrace{P_e[O = o|G = g, A = a, X = x]P_e[G = g|A = a, X = x]}_{\text{Estimable from Case-Control Study}} \underbrace{P_e[A = a|X = x]}_{\text{Estimable from Survey}} \end{aligned}$$

Thus, by using both data sources, $P_e[O = o, G = g|X = x]$ becomes estimable.

For inference, we posit parametric models for $P_e[O = o|G = g, A = a, X = x]$, $P_e[G = g|A = a, X = x]$ and $P_e[A = a|X = x]$. Specifically, we posit a logistic regression model for obesity given genotype, asthma and covariates:

$$\text{logit}\{P_e[O = 1|G = g, A = a, X = x]\} = h(e, g, a, x; \gamma); \tag{2.1}$$

a proportional odds model for genotype given asthma and covariates:

$$\text{logit}\{P_e[G \leq g|A = a, X = x]\} = \beta_{0,g} + \beta_{1,e} + \beta_{2,e}a \quad g = 0, 1 \tag{2.2}$$

and a logistic regression model for asthma given demographic factors:

$$\text{logit}\{P_e[A = 1|X = x]\} = l(e, x; \delta) \tag{2.3}$$

where $h(e, g, a, x; \gamma)$ is a specified function of e, g, a, x and parameter vector γ, there exists one level of e for which $\beta_{1,e} = 0$ and $l(e, x; \delta)$ is a specified function of e, x and parameter vector δ. In model 2.2, we assume that genotype is independent of demographic factors (X) given asthma status; the data do not provide evidence against this assumption. In our analysis, X is `age` and `gender`, and we set, after model fitting,

$$\begin{aligned} h(e, g, a, x; \gamma) \;&=\; \gamma_{0,e} + \gamma_1 a + \gamma_2 I(g = 1) + \gamma_3 I(g = 2) + \gamma_4 \texttt{age} + \gamma_5 \texttt{gender} + \\ &\quad\; \gamma_6 I(g = 1) \cdot a + \gamma_7 I(g = 2) \cdot a + \gamma_8 \texttt{age} \cdot a + \gamma_9 \texttt{gender} \cdot a \end{aligned}$$

$$l(e, x; \delta) = \delta_{0,e} + \delta_2\texttt{gender} + \delta_3\texttt{ns}(\texttt{age}; 5, 11) + \delta_4\texttt{gender} \cdot \texttt{ns}(\texttt{age}; 5, 11)$$

where $\texttt{ns}(\texttt{age}; 5, 11)$ is a B-spline basis for a natural cubic spline with knots at ages 5 years and 11 years. The spline functions were used to model the non-linear dependence of prevalence of asthma with age. The parameters from models 2.1, 2.2 and 2.3 can be estimated using the R functions glm, polr and survglm, respectively. In estimating the parameters of model 2.3, survey weights (obtained from the sample survey) are utilized. The R functions output parameter estimates $\widehat{\gamma}$, $\widehat{\beta}$ and $\widehat{\delta}$ and associated estimated variance-covariance matrices denoted by $\widehat{\Sigma}_{\widehat{\gamma}}$, $\widehat{\Sigma}_{\widehat{\beta}}$, and $\widehat{\Sigma}_{\widehat{\delta}}$, respectively. The parameter estimators from these models are asymptotically normal and asymptotically uncorrelated.

We estimate $P_e[O = o, G = g]$ by Monte Carlo integration using

$$\widehat{P}_e[O = o, G = g] = \frac{1}{M} \sum_{m=1}^{M} \widehat{P}_e[O = o, G = g | X = x_m]$$

where M is a large number; $x_1, \ldots, x_M$ are independent draws from distribution $F(x)$,

$$\begin{aligned} &\widehat{P}_e[O = o, G = g | X = x] \\ &= \sum_{a=0}^{1} \widehat{P}_e[O = o | G = g, A = a, X = x] \widehat{P}_e[G = g | A = a, X = x] \widehat{P}_e[A = a | X = x] \end{aligned}$$

$$\widehat{P}_e[O = o | G = g, A = a, X = x] = \frac{\exp(o \times h(e, g, a, x; \widehat{\gamma}))}{1 + \exp(h(e, g, a, x; \widehat{\gamma}))}$$

$$\widehat{P}_e[G=g|A=a,X=x]=\begin{cases}\frac{\exp(\widehat{\beta}_{0,0}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)}{1+\exp(\widehat{\beta}_{0,0}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)} & g=0\\ \frac{\exp(\widehat{\beta}_{0,1}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)}{1+\exp(\widehat{\beta}_{0,1}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)}-\frac{\exp(\widehat{\beta}_{0,0}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)}{1+\exp(\widehat{\beta}_{0,0}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)} & g=1\\ \frac{1}{1+\exp(\widehat{\beta}_{0,1}+\widehat{\beta}_{1,e}+\widehat{\beta}_{2,e}a)} & g=2\end{cases}$$

$$\widehat{P}_e[A=a|X=x]=\frac{\exp(a\times l(e,x;\widehat{\delta}))}{1+\exp(l(e,x;\widehat{\delta}))}$$

Given the categorical nature of the genotype and phenotype data, there are different ways of expressing their association. One way is to consider ethnicity-specific odds ratios. The three odds ratios we consider are the recessive, dominance and additive odds ratios. The recessive and dominance odds ratios are estimated by

$$\widehat{\text{Recessive Odds}}\text{ Ratio}=\frac{\widehat{P}_e[O=1,G=2]\widehat{P}_e[O=0,G=0,1]}{\widehat{P}_e[O=1,G=0,1]\widehat{P}_e[O=0,G=2]}$$

$$\widehat{\text{Dominance Odds}}\text{ Ratio}=\frac{\widehat{P}_e[O=1,G=1,2]\widehat{P}_e[O=0,G=0]}{\widehat{P}_e[O=1,G=0]\widehat{P}_e[O=0,G=1,2]}$$

We can also estimate the ethnicity-specific additive odds ratio ($\exp(\eta_e)$) from the following model:

$$\text{logit}P_e[O=1|G=g]=\eta_{0,e}+\eta_e g$$

by minimizing (with respect to $\eta_{0,e}$ and η_e)

$$\mathcal{L}(\eta_{0,e},\eta_e)=\sum_{g=0}^{2}\widehat{P}_e[G=g]\left[\widehat{P}_e[O=1|G=g]-\frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right]^2$$

where

$$\begin{aligned}\widehat{P}_e[G=g] &= \sum_{o=0}^{1}\widehat{P}_e[O=o,G=g]\\ \widehat{P}_e[O=1|G=g] &= \frac{\widehat{P}_e[O=1,G=g]}{\widehat{P}_e[G=g]}\end{aligned}$$

This latter estimation procedure is called weighted minimum distance estimation (Klugman and Parsa, 1994).

Since these odds ratio estimators are smooth functions of $\widehat{\gamma}$, $\widehat{\beta}$ and $\widehat{\delta}$, they will also be asymptotically normal. In the Appendix (2.6.1), we present estimates of the standard errors of these estimators. In our analysis, we construct normality-based confidence intervals on the log scale and then exponentiate.

2.4 Results

GALA II study

There were 3757 individuals in the GALA II dataset; 1786, 1245, 105 and 621 were classified as Puerto Rican, Mexican, Mixed Latino and Other Latino, respectively. The age range was 8-21 years. The SNP of interest, *rs12232497*, has major allele T and minor allele C. Among Puerto Ricans, 9 had missing information on the SNP of interest and 618 had missing information on body mass index (BMI), derived from height and weight and used to determine obesity status. Among Mexicans, these numbers were 1 and 171, respectively. There was no missing SNP information among Mixed and Other Latinos. BMI was missing on 9 and 178 among Mixed Latinos and Other Latinos, respectively. The majority of the missing obesity information was among controls. This is because controls were not originally scheduled to be given spirometry tests and these tests require the collection of information on height and weight. There was differential missingness of BMI by ethnicity. This was due to the multi-site nature of the study. Each site had different recruitment goals and ethnic profiles. When the policy to collect height and weight among controls was instituted, sites who recruited more Puerto Ricans and Other Latinos were further along in their recruitment goals than sites who recruited more Mexicans and Mixed Latinos. Our analysis uses data on patients who have completely recorded SNP and BMI, which is 1163 Puerto Ricans (886 cases, 277 controls), 1073 Mexicans (585 cases and 488 controls), 96 Mixed Latinos (61 cases, 35 controls) and 443 Other Latinos (337 cases, 106 controls). The validity of our analysis hinges on the additional, untestable, albeit plausible assumption, that missingness of BMI and SNP data is unrelated to obesity status and the gene given case/control status, ethnicity, age and gender.

Table 2.1 displays various measures of the adjusted (for age and gender) association between asthma and the genotype based on the case-control data for Puerto Ricans and Mexicans. These results suggest that the minor allele C is associated with a decreased risk of asthma in Puerto Ricans (additive odds ratio = 0.66 [95% CI: 0.54, 0.82]) and Mexicans (additive odds ratio = 0.70 [95% CI: 0.57, 0.85]). The

Table 2.1: Measures of marginal association between asthma and genotype adjusted for age and gender based on the GALA II study.

Measure of G-A association	Puerto Ricans Estimate	(95% CI)	Mexicans Estimate	(95% CI)	Interaction Estimate	(95% CI)
Dominance odds ratio	0.60	(0.45,0.79)	0.68	(0.53,0.87)	0.88	(0.61,1.28)
Recessive odds ratio	0.59	(0.37,0.95)	0.52	(0.33,0.82)	1.12	(0.59,2.16)
Additive odds ratio	0.66	(0.54,0.82)	0.70	(0.57,0.85)	0.94	(0.71,1.26)

association is not significantly different between these ethnic subgroups.

Tables 2.2 and 2.3 present the observed frequency distribution of obesity and genotype from the case-control data for Puerto Ricans and Mexicans, respectively. Table 2.4 presents adjusted (for age, gender and asthma status) measures of association between obesity and the genotype for Puerto Ricans and Mexicans, based on the case-control data. Based on this naive analysis, there is no significant association between obesity and genotype (all confidence intervals cover 1).

Tables 2.5 and 2.6 present the results of fitting Models 2.1 and 2.2 based on the case-control data.

NHIS study

The NHIS dataset contains information on 11,277 children in the age range 0-17 years; 167, 311, 2285, 102, 111, 489, 13, 40, 7759 were classified as Multiple Hispanic, Puerto Rican, Mexican, Cuban/Cuban American, Dominican (Republic), Central or South American, Other Latin American (type not specified), Other Spanish, Not

Table 2.2: Observed frequency distribution of obesity and genotype in Puerto Ricans from the GALA II study (percentages shown in parenthesis separately for cases and controls).

Genotype	Cases			Controls		
	Non-obese	Obese	Total	Non-obese	Obese	Total
0	325(36.68)	123(13.88)	448(50.56)	84(30.33)	23(8.30)	107(38.63)
1	264(29.80)	111(12.53)	375(42.33)	107(38.63)	33(11.91)	140(50.54)
2	42(4.74)	21(2.37)	63(7.11)	29(10.47)	1(0.36)	30(10.83)
Total	631(71.22)	255(28.78)	886(100)	220(79.43)	57(20.57)	277(100)

Table 2.3: Observed frequency distribution of obesity and genotype in Mexicans from the GALA II study (percentages shown in paranthesis separately for cases and controls).

Genotype	Cases			Controls		
	Non-obese	Obese	Total	Non-obese	Obese	Total
0	175(29.91)	139(23.76)	314(53.67)	163(33.40)	58(11.89)	221(45.29)
1	152(25.98)	84(14.36)	236(40.34)	136(27.87)	79(16.19)	215(44.06)
2	17(2.91)	18(3.08)	35(5.99)	39(7.99)	13(2.66)	52(10.65)
Total	344(58.80)	241(41.20)	585(100)	338(69.26)	150(30.74)	488(100)

Table 2.4: Measures of marginal association between obesity and genotype in the naive analysis adjusted for age, gender and asthma based on GALA II study.

Measure of G-O association	Puerto Ricans		Mexicans		Interaction	
	Estimate	(95% CI)	Estimate	(95% CI)	Estimate	(95% CI)
Dominance odds ratio	1.05	(0.81,1.37)	0.96	(0.75,1.23)	1.10	(0.76,1.57)
Recessive odds ratio	0.82	(0.49,1.33)	0.96	(0.60,1.50)	0.85	(0.43,1.66)
Additive odds ratio	0.99	(0.81,1.22)	0.97	(0.79,1.17)	1.03	(0.77,1.37)

Table 2.5: Results from obesity model (2.1) based on GALA II study.

Covariate	Estimate	Std. error	Z-value	P
Mexican	-0.77	0.33	-2.31	0.021
Mixed Latino	-0.84	0.40	-2.12	0.034
Other Latino	-0.81	0.35	-2.33	0.020
Puerto Rican	-1.35	0.34	-4.00	< 0.0001
Asthma	0.65	0.39	1.67	0.096
Gender	-0.28	0.15	-1.81	0.070
Age	0.0037	0.022	0.17	0.86
One Copy	0.18	0.16	1.11	0.270
Two Copies	-0.60	0.30	-2.00	0.045
Asthma*Gender	0.013	0.18	0.07	0.95
Asthma*Age	-0.0072	0.027	-0.27	0.79
Asthma*One Copy	-0.31	0.19	-1.65	0.099
Asthma*Two Copies	0.78	0.36	2.17	0.030

Table 2.6: Results from genotype model (2.2) based on GALA II study (Mexican is reference ethnicity).

Covariate	Estimate	Std. Error	Z-value	P
Zero Copies	-0.22	0.09	-2.49	0.013
One Copy	2.27	0.10	21.64	< 0.0001
Mixed Latino	0.10	0.32	0.32	0.748
Other Latino	0.20	0.21	0.98	0.328
Puerto Rican	0.22	0.14	1.55	0.121
Asthma (Mexicans)	-0.38	0.12	-3.19	0.001
Asthma (Mixed Latino)	-0.55	0.40	1.37	0.172
Asthma (Other Latino)	-0.63	0.22	2.92	0.004
Asthma(Puerto Ricans)	-0.47	0.13	3.59	0.0003

Hispanic/Spanish origin respectively. 21 children have missing information on asthma status; 3 of them were Mexican, 1 Central or South American and 17 were not of Hispanic/Spanish origin. Of the 21 children with missing asthma status, 4 refused response and 17 did not know their asthma status information. Estimation of Model 2.3 made use of data on 11,256 children (1569 with asthma and 9687 without asthma) who have complete information on asthma status. Table 2.7 presents the results of fitting Model 2.3 using the survey weights.

Figure 2.1 shows the variation of prevalence of asthma with age for different genders in Puerto Ricans and Mexicans as explained by Model 2.3. We observe a steep increase in the prevalence of asthma with age in the range 0-7 years; Puerto Ricans have a greater rate of this increase compared to Mexicans. The relationship gets flatter for older males in both ethnic groups. The younger females have a lower prevalence of asthma compared to younger males. After a steep increase in the age range 0-7

Table 2.7: Results from asthma model (2.3) based on NHIS 2010.

Covariate	Estimate	Std. Error	t-value	P
Multiple Hispanic	-2.9496	0.3408	-8.66	< 0.0001
Puerto Rican	-1.9454	0.2495	-7.80	< 0.0001
Mexican	-3.0946	0.2191	-14.12	< 0.0001
Cuban/Cuban American	-2.9441	0.5053	-5.83	< 0.0001
Dominican (Republic)	-2.6618	0.3690	-7.21	< 0.0001
Central or South American	-3.1891	0.2527	-12.62	< 0.0001
Other Latin American, type not specified	-3.9114	1.0683	-3.66	0.0003
Other Spanish	-3.6461	0.6467	-5.64	< 0.0001
Not Hispanic/Spanish origin	-2.8233	0.2053	-13.75	< 0.0001
Gender	-0.4815	0.3197	-1.51	0.1332
ns(age; 5,11)1	0.9582	0.1895	5.06	< 0.0001
ns(age; 5,11)2	2.5305	0.4590	5.51	< 0.0001
ns(age; 5,11)3	0.5179	0.1399	3.70	0.0003
Gender × ns(age; 5,11)1	0.0926	0.2993	0.31	0.7572
Gender × ns(age; 5,11)3	0.3806	0.7110	0.54	0.5929
Gender × ns(age; 5,11)3	0.4418	0.1928	2.29	0.0227

Table 2.8: Measures of marginal association between obesity and genotype from the combined analysis based on GALA II Study and NHIS 2010.

Measure of G-O association	Puerto Ricans		Mexicans		Interaction	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Dominance odds ratio	0.95	0.78, 1.17	1.01	0.79, 1.30	0.94	0.86, 1.03
Recessive odds ratio	0.65	0.44, 0.95	0.56	0.36, 0.89	1.15	0.94, 1.41
Additive odds ratio	0.90	0.78, 1.04	0.91	0.77, 1.09	0.98	0.92, 1.05

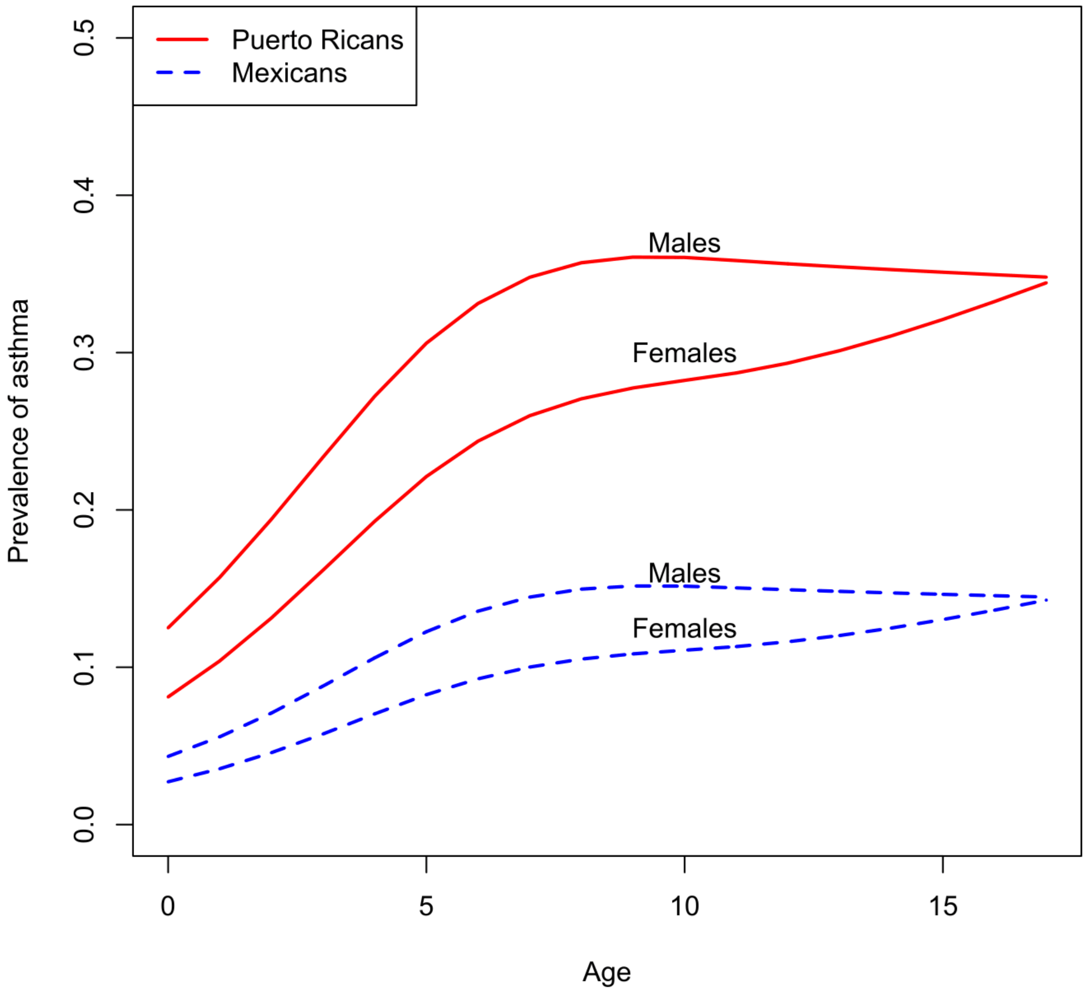


Figure 2.1: Variation of asthma prevalence with age (in years), gender and ethnicity as explained by Model (2.3). Note that we used B-spline basis for a natural cubic spline to smooth over age.

years, the prevalence of asthma for females keeps increasing but at a slower rate. In the older age-groups, the differences in the prevalence of asthma between males and females decreases. Thus, there is evidence of interaction between age and gender.

Combined analysis

In our combined analysis we worked with a reference population of size M = 2000. Table 2.8 presents our results. The recessive odds ratio between obesity and genotype is significantly less than 1 in both Puerto Ricans and Mexicans, i.e. the individuals with 2 copies of the minor allele are at a lower risk of obesity compared to 0 or 1 copy.

Figure 2.2 shows the point estimates and 95% confidence intervals for log-odds of $P_e[O = 1|G = g]$ for $g = 0, 1, 2$ for both Puerto Ricans and Mexicans. The plots show evidence in favor of a recessive inheritance model. For each ethnic group, the points in Figure 2.2 are not on a straight line; a strong indication that the additivity assumption may not hold. We computed point estimates and 95% confidence intervals for $\tau_e = \text{logit}(P_e[O = 1|G = 2]) - 2\text{logit}(P_e[O = 1|G = 1]) + \text{logit}(P_e[O = 1|G = 0])$ for both Puerto Ricans and Mexicans. When the additivity assumption holds $\tau_e = 0$. For Mexicans, the evidence against additivity assumption is statistically significant [$\hat{\tau}_e = -0.73$, 95% CI: $-1.34, -0.13$]; for Puerto Ricans it is of borderline significance [$\hat{\tau}_e = -0.46$, 95% CI: $-0.98, 0.05$].

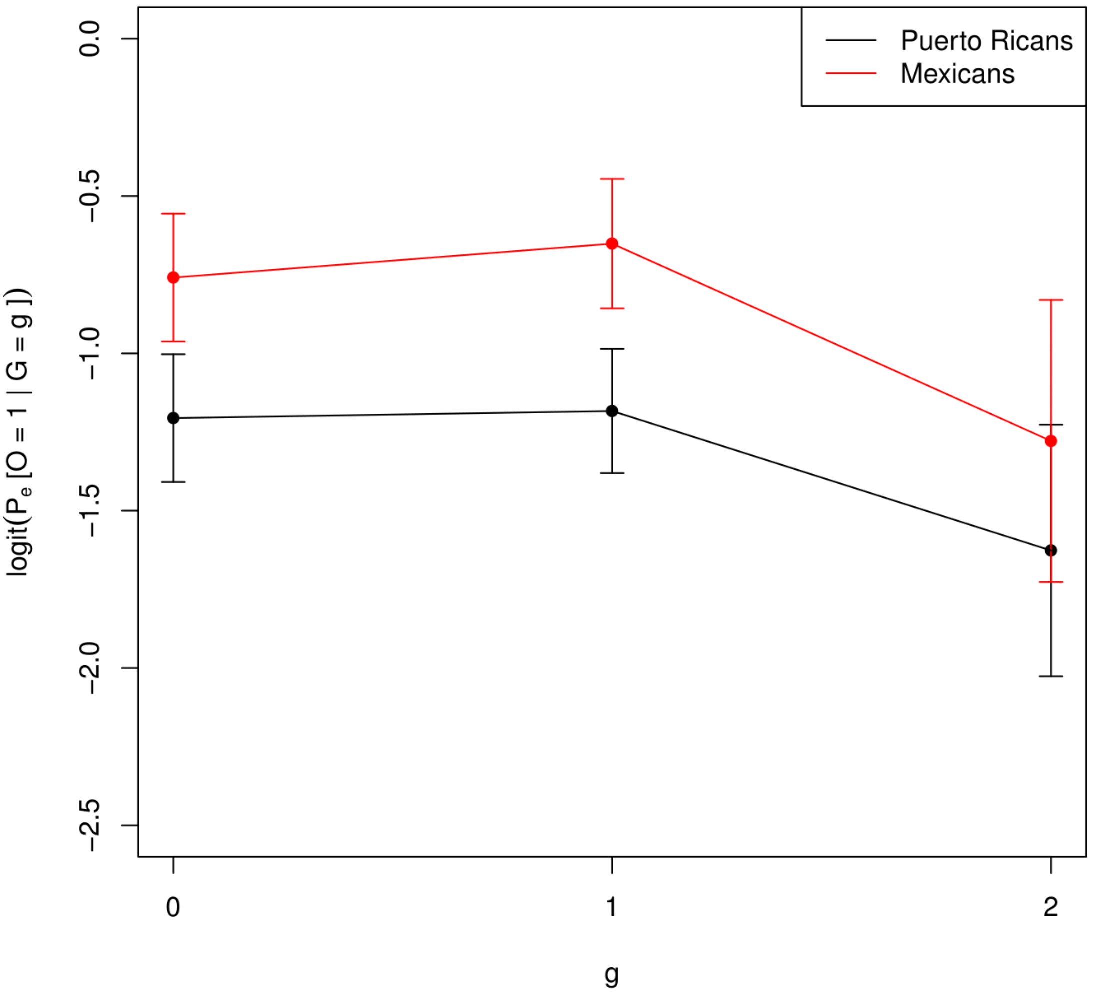


Figure 2.2: Variation of log-odds of obesity with genotype in Puerto Ricans and Mexicans obtained from our methodology. We show the point estimates and 95% confidence intervals (vertical bars) for the log-odds of obesity at the different levels of the gene (i.e. logit$(Pe[O = 1|G = g])$ for $g = 0, 1, 2$) for both Puerto Ricans and Mexicans.

2.5 Discussion

The analysis of secondary phenotypes in genetic case-control studies are subject to bias. We presented an approach to mitigate this bias by integrating information from representative sample surveys. In the combined analysis we found that individuals with 2 copies of the minor allele are at a lower risk of obesity compared to 0 or 1 copy. The naive analysis of the GALA II dataset that ignores the selective sampling of cases and controls results in null findings. This illustrates the drawbacks of the naive analysis of case-control data. Our statistical framework for estimating uncertainty includes sampling uncertainty from both the sample survey and the case-control study. More generally, our framework allows one to obtain population-level estimates of genetic effects on clinical quantities (e.g., serum glucose level, concentration of a metabolite etc) that would be hard to measure in a large scale sample survey.

The conditional independence assumptions of Nagelkerke *and others* (1995) and Kraft (2007) do not hold in our case. In particular, there is statistical evidence that asthma status is not conditionally independent of the genotype given obesity status for Puerto Ricans (Cochran-Mantel-Haenszel P = 0.001) and Mexicans (Cochran-Mantel-Haenszel P = 0.003); and asthma status is not conditionally independent of obesity status given genotype for Puerto Ricans (Mantel-Haenszel P = 0.008) and Mexicans (Mantel-Haenszel P = 0.0005).

In our setting, the sampling fractions of cases and controls are not known. Using an approach that requires specification of the prevalence of asthma is difficult because it is essential that it be computed from a population where the conditional distribution of key risk factors given asthma status matches that in the case-control study. Furthermore, it is important to reflect the uncertainty associated with the estimate of prevalence. It is possible to show theoretically that when we do not control for demographic factors (e.g. age and gender) the 95% confidence intervals for $P[O = o, G = g]$ will be wider when prevalence of asthma is estimated with uncertainty from an external data source (e.g. sample survey) relative to when it is assumed known (details in Appendix (2.6.2)). In our example this increase in width is small but consistent with theory (data not shown).

Since asthma is a common disease, the rare disease assumption by Li *and others* (2010) is not justified. The profile likelihood method of Lin and Zeng (2009) profiles out the distribution of the genetic risk factor when the disease is common and prevalence of disease is unknown. We implemented their method for our case-control dataset, but the model parameters (including the prevalence of asthma) are not identifiable in the sense that multiple maximizers of their profile likelihood were found. Our proposed methodology does not make the above assumptions and also provides a framework for control of key demographic factors.

In our framework, the analyst has to choose the structure of the reference population. We chose a population with equal sex ratios and a uniform age distribution. In general, there may be disagreement on the appropriate reference population, but sensitivity to such disagreement is easily examined by considering a range of reference population characteristics.

The use of our method assumes the existence of a sample survey where the primary phenotype of the case-control study is also measured. There may be differences in how the phenotype is measured in the two data sources. This can lead to additional biases. For example, in GALA II, measurement of asthma was based upon physician diagnosis. In contrast, the NHIS survey used self-report from an adult in the household about whether the child had a physician diagnosis of asthma. Moreover, in the case-control study, the individuals within asthma-age-gender-ethnicity strata may not be representative. This can also lead to some bias.

In general, the covariates such as age and gender should be independent of genotype in the population. Adding this constraint can lead to efficiency improvement. Similarly, it may be reasonable to assume the genotypes are in Hardy-Weinberg equilibrium in the population. This constraint could also lead to improvements in efficiency. These arguments for improving efficiency rely on the modeling assumptions being correctly specified and if not, they might introduce bias. Thus the analyst has to make careful choices in trading off bias and variance.

Genetic case-control studies typically characterize subjects in great clinical detail, making it difficult to conduct on a large scale. Moreover, these studies are biased by design. Sample surveys are designed to be representative, but do not allow detailed clinical characterization. Our method provides a statistical framework to leverage the strengths of sample surveys with case-control studies to provide unbiased genetic association estimates of clinical phenotypes that are hard to measure in large scale surveys.

2.6 Appendix

2.6.1 Computation of standard errors

In Section 2.3 we saw that for a particular ethnicity stratum e, the 6×1 vector of probabilities $\{P_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ can be expressed as a 6-variate smooth function $f(\gamma^*, \beta^*, \delta^*)$, where $\gamma^*, \beta^*, \delta^*$ are the true values of the parameters γ, β, δ respectively. The parameter estimates $\hat{\gamma}$, $\hat{\beta}$ and $\hat{\delta}$ and their estimated variance-covariance matrices $\hat{\Sigma}_{\hat{\gamma}}$, $\hat{\Sigma}_{\hat{\beta}}$ and $\hat{\Sigma}_{\hat{\delta}}$ are obtained by fitting the models 2.1, 2.2 and 2.3 in Section 2.3. The parameter estimates are asymptotically normal and asymptotically uncorrelated. A simple application of Multivariate Delta Theorem shows that the distribution of the centered and scaled vector of probabilities $\widehat{P}_e[O = o, G = g]$ is asymptotically normal with variance covariance matrix given by $D\Sigma D'$ where D is the appropriate matrix of derivatives and Σ is the block diagonal matrix with the blocks given by $\Sigma_{\hat{\gamma}}$, $\Sigma_{\hat{\beta}}$ and $\Sigma_{\hat{\delta}}$ respectively. The ethnicity specific dominance odds ratio and recessive odds ratio defined in Section 2.3 are smooth functions of $\{P_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ and hence another application of Delta Theorem gives us the asymptotic distribution of the point estimates of these association measures.

Note that the ethnicity specific additive odds ratio ($exp(\eta_e)$) is obtained from the following model:

$$\text{logit} P_e[O=1|G=g] = \eta_{0,e} + \eta_e g$$

by minimizing (with respect to $\eta_{0,e}$ and η_e)

$$\mathcal{L}(\eta_{0,e},\eta_e) = \sum_{g=0}^{2} \widehat{P}_e[G=g]\left[\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right]^2$$

where

$$\begin{aligned} \widehat{P}_e[G=g] &= \sum_{o=0}^{1} \widehat{P}_e[O=o,G=g] \\ \widehat{P}_e[O=1|G=g] &= \frac{\widehat{P}_e[O=1,G=g]}{\widehat{P}_e[G=g]} \end{aligned}$$

This latter estimation procedure is called weighted minimum distance estimation. In what follows we discuss how to compute standard errors of the ethnicity specific additive odds ratio. Note that minimization of $\mathcal{L}(\eta_{0,e},\eta_e)$ is equivalent to solving the system of equations:

$$\frac{\partial \mathcal{L}}{\partial \eta_{0,e}} = \sum_{g=0}^{2} \widehat{P}_e[G=g]\frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2}\left[\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right] = 0 \tag{2.4}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_{e}} = \sum_{g=0}^{2} g\widehat{P}_e[G=g]\frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2}\left[\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right] = 0 \tag{2.5}$$

Denote the solution by $(\widehat{\eta}_{0,e},\widehat{\eta}_e)$. A first order Taylor series expansion around $(\eta_{0,e},\eta_e)$ of 2.4 is given by:

$$\begin{aligned} 0 &= \sum_{g=0}^{2} \widehat{P}_e[G=g]\frac{e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}{(1+e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g})^2}\left[\widehat{P}_e[O=1|G=g]-\frac{e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}{1+e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}\right] \\ &= \sum_{g=0}^{2} \widehat{P}_e[G=g]\frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2}\left[\widehat{P}_e[O=1|G=g]-\frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right] \\ &\quad + \begin{pmatrix} \frac{\partial^2\mathcal{L}}{\partial\eta_{0,e}^2}(\eta_{0,e}^{(1)},\eta_e^{(1)}) & \frac{\partial^2\mathcal{L}}{\partial\eta_e\partial\eta_{0,e}}(\eta_{0,e}^{(1)},\eta_e^{(1)}) \end{pmatrix} \begin{pmatrix} \widehat{\eta}_{0,e}-\eta_{0,e} \\ \widehat{\eta}_e-\eta_e \end{pmatrix} \end{aligned} \tag{2.6}$$

where $||(\eta_{0,e}^{(1)},\eta_e^{(1)})-(\eta_{0,e},\eta_e)|| < ||(\widehat{\eta}_{0,e},\widehat{\eta}_e)-(\eta_{0,e},\eta_e)||$. A similar expansion of 2.5 gives us:

$$\begin{aligned} 0 &= \sum_{g=0}^{2} g\widehat{P}_e[G=g]\frac{e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}{(1+e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g})^2}\left[\widehat{P}_e[O=1|G=g]-\frac{e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}{1+e^{\widehat{\eta}_{0,e}+\widehat{\eta}_e g}}\right] \\ &= \sum_{g=0}^{2} g\widehat{P}_e[G=g]\frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2}\left[\widehat{P}_e[O=1|G=g]-\frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}}\right] \\ &\quad + \begin{pmatrix} \frac{\partial^2\mathcal{L}}{\partial\eta_e\partial\eta_{0,e}}(\eta_{0,e}^{(2)},\eta_e^{(2)}) & \frac{\partial^2\mathcal{L}}{\partial\eta_e^2}(\eta_{0,e}^{(2)},\eta_e^{(2)}) \end{pmatrix} \begin{pmatrix} \widehat{\eta}_{0,e}-\eta_{0,e} \\ \widehat{\eta}_e-\eta_e \end{pmatrix} \end{aligned} \tag{2.7}$$

where $||(\eta_{0,e}^{(2)},\eta_e^{(2)})-(\eta_{0,e},\eta_e)|| < ||(\widehat{\eta}_{0,e},\widehat{\eta}_e)-(\eta_{0,e},\eta_e)||$. The expressions for the double partial derivatives of $\mathcal{L}(.,.)$ in 2.6 and 2.7 are given by:

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_{0,e}^2}(\eta_{0,e}^{(1)}, \eta_e^{(1)}) = \sum_{g=0}^{2} \widehat{P}_e[G=g]\Bigg[\frac{e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}{\left(1+e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}\right)^4}\Bigg\{(1-e^{2\eta_{0,e}^{(1)}+2\eta_e^{(1)}g}) \times \left(\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}{1+e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}\right) - e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}\Bigg\}\Bigg]$$

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_e \partial \eta_{0,e}}(\eta_{0,e}^{(1)}, \eta_e^{(1)}) = \sum_{g=0}^{2} g\widehat{P}_e[G=g]\Bigg[\frac{e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}{\left(1+e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}\right)^4}\Bigg\{(1-e^{2\eta_{0,e}^{(1)}+2\eta_e^{(1)}g}) \times \left(\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}{1+e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}}\right) - e^{\eta_{0,e}^{(1)}+\eta_e^{(1)}g}\Bigg\}\Bigg]$$

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_e \partial \eta_{0,e}}(\eta_{0,e}^{(2)}, \eta_e^{(2)}) = \sum_{g=0}^{2} g\widehat{P}_e[G=g]\Bigg[\frac{e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}{\left(1+e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}\right)^4}\Bigg\{(1-e^{2\eta_{0,e}^{(2)}+2\eta_e^{(2)}g}) \times \left(\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}{1+e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}\right) - e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}\Bigg\}\Bigg]$$

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_e^2}(\eta_{0,e}^{(2)}, \eta_e^{(2)}) = \sum_{g=0}^{2} g^2\widehat{P}_e[G=g]\Bigg[\frac{e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}{\left(1+e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}\right)^4}\Bigg\{(1-e^{2\eta_{0,e}^{(2)}+2\eta_e^{(2)}g}) \times \left(\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}{1+e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}}\right) - e^{\eta_{0,e}^{(2)}+\eta_e^{(2)}g}\Bigg\}\Bigg]$$

Let $J = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \eta_{0,e}^2}(\eta_{0,e}^{(1)}, \eta_e^{(1)}) & \frac{\partial^2 \mathcal{L}}{\partial \eta_e \partial \eta_{0,e}}(\eta_{0,e}^{(1)}, \eta_e^{(1)}) \\ \frac{\partial^2 \mathcal{L}}{\partial \eta_e \partial \eta_{0,e}}(\eta_{0,e}^{(2)}, \eta_e^{(2)}) & \frac{\partial^2 \mathcal{L}}{\partial \eta_e^2}(\eta_{0,e}^{(2)}, \eta_e^{(2)}) \end{pmatrix}$ and

$$b = \begin{pmatrix} \sum_{g=0}^{2} \widehat{P}_e[G=g] \frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2} \left[\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}} \right] \\ \sum_{g=0}^{2} g\widehat{P}_e[G=g] \frac{e^{\eta_{0,e}+\eta_e g}}{(1+e^{\eta_{0,e}+\eta_e g})^2} \left[\widehat{P}_e[O=1|G=g] - \frac{e^{\eta_{0,e}+\eta_e g}}{1+e^{\eta_{0,e}+\eta_e g}} \right] \end{pmatrix}$$

Solving 2.6 and 2.7, we have:

$$\begin{pmatrix} \widehat{\eta}_{0,e} - \eta_{0,e} \\ \\ \widehat{\eta}_e - \eta_e \end{pmatrix} = -J^{-1}b \tag{2.8}$$

From the asympotic distribution of $\{\widehat{P}_e[O=o,G=g] : o=0,1; g=0,1,2\}$ we derive the asymptotic distributiom of $\{\widehat{P}_e[G=g], \widehat{P}_e[O=1|G=g] : g=0,1,2\}$ by an application of Multivariate Delta Theorem. Similarly from the asymptotic distribution of $\{\widehat{P}_e[G=g], \widehat{P}_e[O=1|G=g] : g=0,1,2\}$, we compute the asymptotic distribution of b. Note that J^{-1} will converge in probability to the inverse of a matrix whose entries are the double partial derivatives of $\mathcal{L}$ evaluated at the true values of the arguments. An application of Slutsky's Theorem in 2.8 gives us the asymptotic distribution of LHS of 2.8 (appropriately scaled) and hence the asymptotic distribution of the additive odds ratio.

2.6.2 Related Asymptotics

Focus on a particular ethnicity stratum $E=e$. Consider the case when we do not control for demographic factors (e.g., age and gender). We want to compare the asymptotic standard errors of $\{\widehat{P}_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ when $P_e[A = 1]$ is known versus when it is estimated with uncertainty from some external data source. First consider the case when it is estimated from an external data source. Note that:

$$P_e[O = o, G = g] = \sum_{a=0}^{1} P_e[O = o, G = g|A = a]P_e[A = a]$$

Note that $\{\widehat{P}_e[O = o, G = g|A = 1] : o = 0, 1; g = 0, 1, 2\}$, $\{\widehat{P}_e[O = o, G = g|A = 0] : o = 0, 1; g = 0, 1, 2\}$ and $\widehat{P}_e[A = 1]$ are asymptotically uncorrelated; denote the asymptotic variances by $\Sigma_{A=1}$, $\Sigma_{A=0}$ and σ_A^2 respectively. The combined variance covariance matrix Σ is block diagonal with $\Sigma_{A=1}$, $\Sigma_{A=0}$ and σ_A^2 as the diagonal blocks. By Multivariate Delta Theorem, the asymptotic variance-covariance matrix for $\{\hat{P}_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ is given by $(P_e[A = 1])^2\Sigma_{A=1} + (1 - P_e[A = 1])^2\Sigma_{A=0} + \sigma_A^2 vv^T$, where v is the vector of the differences $\{P[O = o, G = g|A = 1] - P[O = o, G = g|A = 0] : o = 0, 1; g = 0, 1, 2\}$. Note that $\sigma_A^2 vv^T$ is a non-negative definite matrix. When the prevalence of asthma is assumed known, the last term in the variance expression is not there. This implies, the difference in the variance of $\{\widehat{P}_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ when $P_e[A = 1]$ is estimated with uncertainty versus when it is treated as a known constant is non-negative definite. Hence we have wider 95% confidence intervals for $\{P_e[O = o, G = g] : o = 0, 1; g = 0, 1, 2\}$ when $P_e[A = 1]$ is estimated versus when it is known.

Chapter 3

Causal Effect Among The Exposed: Multiple Data Sources and Censored Outcomes

3.1 Introduction

Consider a setting where a group of autonomous individuals choose to expose themselves to an intervention with potentially adverse consequences. To understand the risk associated with their choice, researchers may be interested in contrasting the distribution of their outcomes under exposure to the intervention to the distribution of their corresponding outcomes had they, contrary to fact, not exposed themselves to the intervention. That is, researchers would like to draw inference about the causal effect among the exposed. Geneletti and Dawid (2011) refer to this estimand as the "effect of treatment on the treated". Our interest in this estimand is motivated by the question of whether individuals who choose to donate kidneys are putting themselves at increased risk for adverse health outcomes such as diabetes and hypertension. Specifically, we would like to learn whether kidney donation accelerates the development of these outcomes.

Since it is not possible to observe the counterfactual outcomes among the exposed individuals, it is necessary to (1) utilize data from non-exposed individuals and (2) posit untestable assumptions in order to learn about the causal effect of interest. In addressing the kidney donation question, we use information available on live kidney donors derived from hospital records and follow-up interviews and on healthy non-donors from two prospective cohort studies. We consider the endpoints of hypertension-free and diabetes-free survival. Our analysis is complicated by the fact that, in the data sources, the endpoint is censored in the broadest sense (i.e., a combination of interval-censored, right censored and exact observations).

In Section 3.2, we develop a method for drawing inference about the causal effect among the exposed based on censored survival outcome data obtained for exposed and non-exposed individuals from different data sources. Section 3.3 applies this method to address our motivating question. Section 3.4 presents a detailed simulation study to evaluate the performance of our methodology. The final section 3.5 is devoted to a discussion.

3.2 Methods

Consider a hypothetical study design in which eligible patients are enrolled and given the option to select "exposure" (e.g., kidney donation) or "non-exposure". Further, assume that the mechanism of exposure selection only depends on observed covariates at the time of enrollment. The patients are then followed from enrollment to the minimum of death or some disease of interest (e.g., hypertension or diabetes). Let Z denote the indicator that the patient opts for "exposure" and W denote the covariates measured at enrollment. Let T_1 and T_0 denote the time from enrollment to death or disease (whichever occurs earlier) for a patient under "exposure" and "non-exposure" respectively. Our goal is to learn about the causal effect among exposed subjects. That is, we want to compare $S_1(t) \overset{def}{=} P[T_1 > t|Z = 1]$ and $S_0(t) \overset{def}{=} P[T_0 > t|Z = 1]$, for all t. The study design is assumed to provide information about the joint distribution of (W, Z, T), where $T = ZT_1 + (1 - Z)T_0$.

3.2.1 Identification of Causal Parameters

In this hypothetical study design we assume

$$Z \perp (T_1, T_0)|W \tag{3.1}$$

i.e. exposure selection depends only on the measured covariates at enrollment. Under Assumption (3.1),

$$S_1(t) = P[T > t|Z = 1] \tag{3.2}$$

and

$$S_0(t) = \int_w P[T > t|W = w, Z = 0]dF(w|Z = 1) \tag{3.3}$$

Under Assumption (3.1), Equations (3.2) and (3.3) provide identification formulae for $S_1(t)$ and $S_0(t)$. From these equations, it follows that in order to estimate $S_1(t)$ and $S_0(t)$, we need to be able to estimate (1) the distribution of T given $Z = 1$, (2) the distribution of W given $Z = 1$, and (3) the distribution of T given W and $Z = 0$.

3.2.2 Manufactured Dataset

Unfortunately, in our setting, it is not possible to conduct the hypothetical study. Rather, we have access to multiple data sources, which we will use to construct a dataset D^* that mimics what might arise from our hypothetical study. To illustrate this construction we will use minimum of hypertension or death as the event of interest. Assume, for the moment, that our data sources provide access to exact times of the event of interest.

Our first data source, D_1, includes patients who donated kidneys. For these patients, the time of enrollment is the time of kidney donation. Figure 3.1(a) displays three patients, numbered 1, 2, 3 in green, from this data source. Patient 1 donates a kidney in 1975, develops hypertension in 1990 (black cross) and dies in 2000 (black asterisk). His time to event is time since kidney donation to the development of hypertenstion (i.e., 15 years). Patient 2 donates a kidney in 1990 and dies without developing hypertension in 2010. His time to event is time since kidney donation to death (i.e., 20 years). Patient 3 donates a kidney in 1980 and dies without developing hypertension in 1995. His time to event is time since kidney donation to death (i.e., 15 years). In contrast to the next data source, all of these patients have a single enrollment visit, denoted by v_1 in the figure.

Our second data source, D_0, includes patients who have not donated kidneys. Each patient has possibly multiple examination times (i.e., multiple visits $v_1, v_2, \ldots$ marked with blue dots). Figure 3.1(c) shows three patients, numbered 1, 2, 3 in blue, from this data source. Patient 1 enters the study in 1985 (marked with label v_1), has a follow-up examination in 1988 (marked with label v_2) and eventually dies without developing hypertension in 1998. Patient 2 enters the study in 1988 and has three follow-up examinations in 1991, 1995 and 2009. He develops hypertension in 2005 (between the third and fourth follow-up examinations) and eventually dies in 2012. Patient 3 enters the study in 1991, develops hypertension in 2001 and dies in 2005.

Figure 3.1(b) is the manufactured dataset D^* that represents (on a study time scale) our hypothetical study described above. This dataset is created by patching together D_1 and D_0 as follows. Each patient in D_1 contributes one enrollment to the hypothetical dataset, i.e. patients 1, 2 and 3 in Figure 3.1(a) contributes the first three enrollments in Figure 3.1(b). Each patient in the data source D_0 (cf Figure 3.1(c)) contributes an "enrollment" at each examination time that occurs before the event the of interest. That is, time of each "enrollment" is considered as a potential time at which the patient could have been eligible to donate a kidney. Patient 1 contributes two "enrollments" to the hypothetical dataset (i.e. the fourth and fifth "enrollments" in Figure 3.1(b)). Patient 2 contributes three "enrollments" to the hypothetical dataset (i.e. sixth, seventh and eighth "enrollments" in Figure 3.1(b)). The fourth visit of Patient 2 is not considered an "enrollment" since it occurs after the event. Patient 3 contributes the last "enrollment" in Figure 3.1(b). Note that in both the data sources we have measured covariates W (age, gender, race, BMI) on the patients at each "enrollment". The idea of multiple enrollments for individual patients was employed by Hernán *and others* (2005) to estimate the causal effect of a time varying exposure on a possibly right censored survival outcome.

If we can think of this manufactured dataset as representing our hypothetical study design, then we are able to identify the causal parameters via equations (3.2) and (3.3). This includes making the working assumption that all entries into the manufactured dataset are independent. We relax this assumption when characterizing the uncertainty of our estimation procedure.

In reality, we do not observe exact times to event in D_1 and D_0. Instead what we observe is censored survival data, which is a combination of interval-censored, right censored and exact observations. Figure 3.2 illustrates different censoring scenarios.

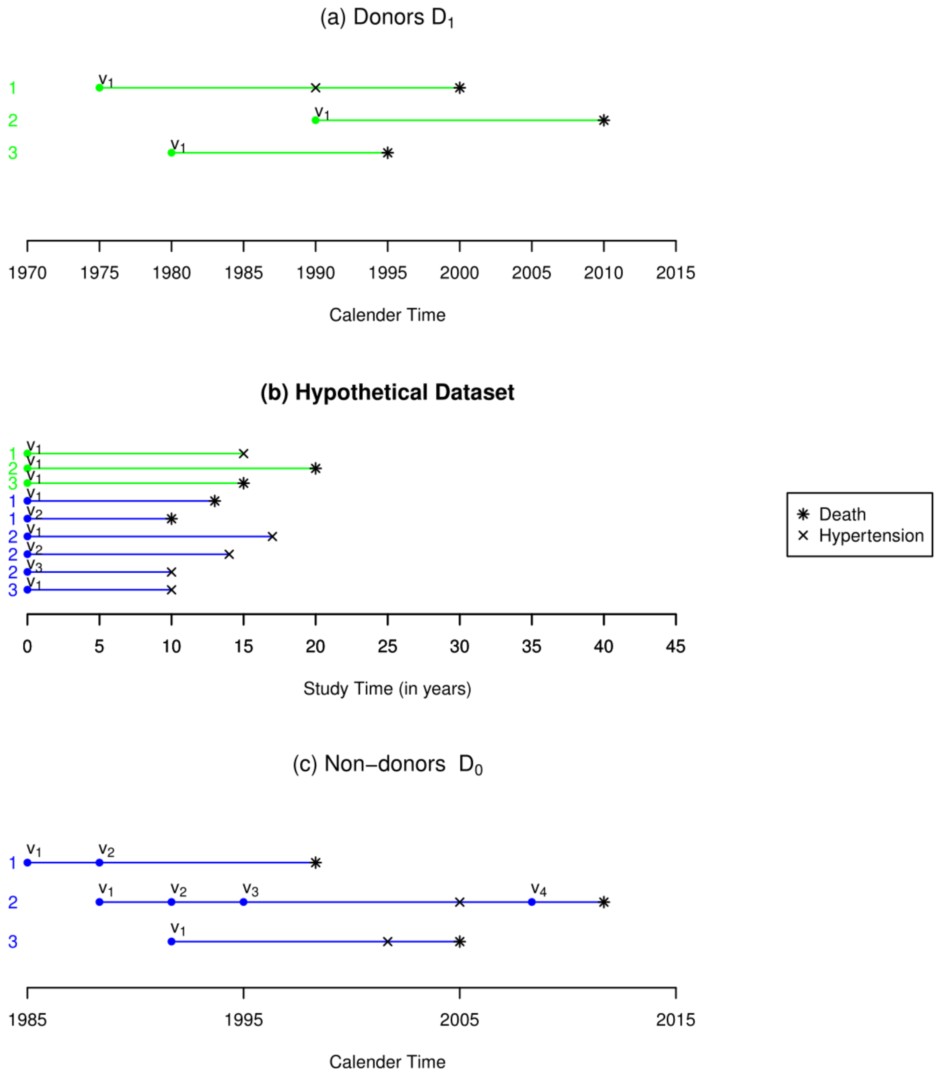


Figure 3.1: Illustration of the process of manufacturing a dataset having the same features as the hypothetical study by patching together two data sources: D_1 from the patients (numbered in green color) who have donated kidneys; and D_0 from the patients (numbered in blue color) who are "eligible" donors but have not donated kidneys. (a) schematic representation of the patients in D_1 with solid green dots denoting "enrollment" (i.e., kidney donation) and green line denoting time from enrollment to either hypertension (black cross) or death (black asterisk); (b) schematic representation of the manufactured hypothetical dataset, (c) schematic representation of the patients in D_0 with solid blue dots denoting multiple "enrollments" and blue line denoting time from an enrollment to either hypertension or death.

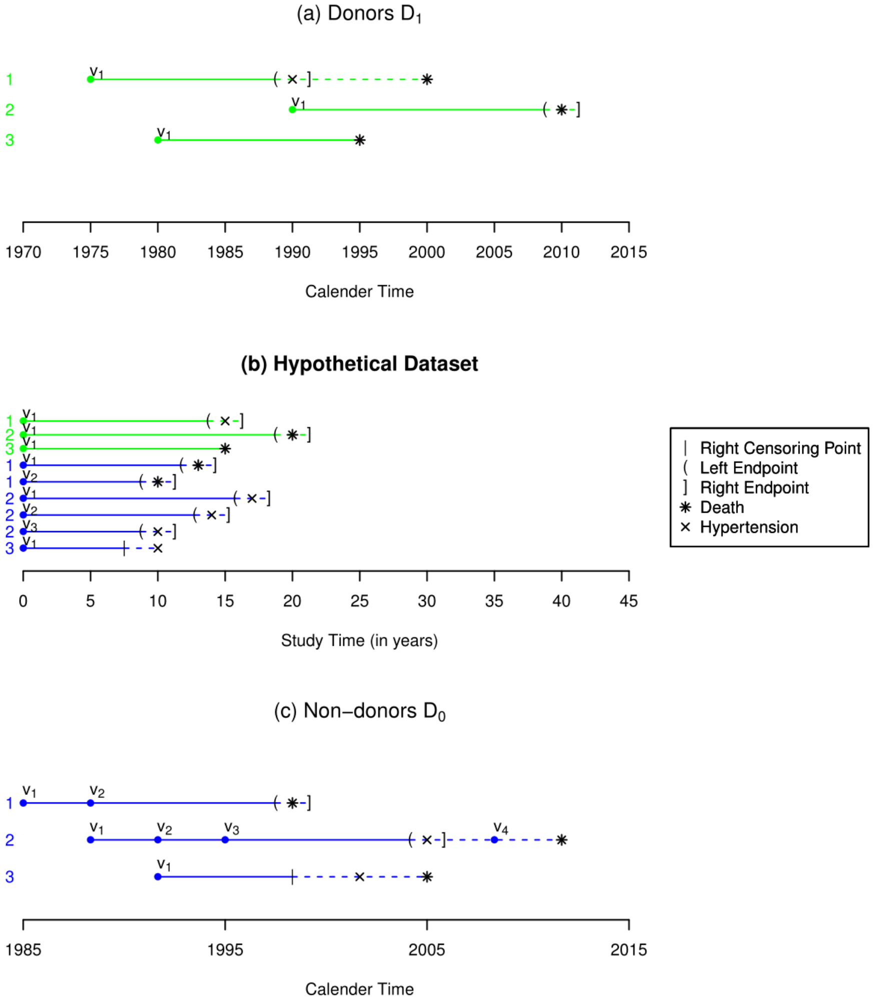


Figure 3.2: Illustration of the process of manufacturing a dataset having the same features as the hypothetical study by patching together two data sources: D_1 from the patients (numbered in green color) who have donated kidneys; and D_0 from the patients (numbered in blue color) who are "eligible" donors but have not donated kidneys. The outcome could be censored (combination of interval-censored, right censored and exact observations); the solid lines become dotted eventually to illustrate the idea of censoring i.e., the exact time of event is not known. (a) schematic representation of the patients in D_1 with solid green dots denoting "enrollment" (i.e., kidney donation) and green line (first solid and then dotted) denoting time from enrollment to either hypertension (black cross) or death (black asterisk); (b) schematic representation of the manufactured hypothetical dataset, (c) schematic representation of the patients in D_0 with solid blue dots denoting multiple "enrollments" and blue line (first solid and then dotted) denoting time from an enrollment to either hypertension or death.

Figure 3.2(a) shows the same three patients as in Figure 3.1(a). Patients 1 and 2 have their times to event interval censored. Patient 3, however, has an exact time of death recorded. Figure 3.2(c) shows the same three patients as in Figure 3.1(c). Patients 1 and 2 have interval censored observations. The outcome for Patient 3 is right-censored (marked by a black vertical bar). Figure 3.2(b) shows the manufactured hypothetical dataset D^* with censored observations.

In the presence of coarsening, additional assumptions are required to identify $P[T > t|Z = 1]$ and the distribution of $P[T > t|W, Z = 0]$. We will assume non-informative censoring conditional Z and W (Gómez *and others*, 2004; Oller *and others*, 2004).

3.2.3 Inference

Our inferential framework aims to contrast $S_1(t)$ and $S_0(t)$, under assumptions, by using Equations (3.2) and (3.3) applied to the manufactured dataset D^*. The key idea is to estimate $P[T > t|Z = 1]$ and $F(w|Z = 1)$ from the donors in D^* and $P[T > t|W = w, Z = 0]$ from the non-donors in D^*. Let $S_1(t|w) \overset{def}{=} P[T > t|W = w, Z = 1]$ and $S_0(t|w) \overset{def}{=} P[T > t|W = w, Z = 0]$. Let n be the number of “enrollments” in D^*. The observed data for each “enrollment” i in D^* is $[E_i, \{T_i : E_i = 1\}, \{(L_i, R_i] : E_i = 0\}, Z_i, W_i]$, where E_i denotes the indicator of exactly observing the failure time, T_i denotes the failure time observed when $E_i = 1$, and L_i and R_i denote the left and right endpoints of the interval in which the time to event is known to lie when

$E_i = 0$. For right censored observations, $L_i < \infty, R_i = \infty$ and for interval-censored observations $L_i < R_i < \infty$.

Under non-informative censoring and independence of "enrollments" in D^*, the simplified likelihood for the observed data (Gómez *and others*, 2004) can be approximated by:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} [\{S_1(L_i|W_i) - S_1(R_i|W_i)\}^{1-E_i}\{(S_1(T_i|W_i) - S_1(T_i+\epsilon|W_i))/\epsilon\}^{E_i}]^{Z_i} \\
&\qquad [\{S_0(L_i|W_i) - S_0(R_i|W_i)\}^{1-E_i}\{(S_0(T_i|W_i) - S_0(T_i+\epsilon|W_i))/\epsilon\}^{E_i}]^{1-Z_i}
\end{aligned}
\tag{3.4}
$$

where ϵ is a specified constant. Note that the "enrollments" with exact observations contribute to the likelihood using a numerical approximation to the conditional densities of T given W and $Z = 1$ and of T given W and $Z = 0$. The numerical approximation is based on the negative of the numerical derivative of the respective survival functions. The numerical derivatives involve the perturbation parameter ϵ which we recommend setting to a small value relative to range of the survival times.

We assume a proportional hazards model (Cox, 1972) for $S_1(t|W)$ and $S_0(t|W)$. Specifically, we assume (for $z = 0, 1$)

$$S_z(t|W) = \exp\{-\Lambda_{0,z}(t)\exp(W'\beta_z)\} \tag{3.5}$$

where β_z is the vector of regression parameters corresponding to the vector of covariates W and $\Lambda_{0,z}(t)$ is the cumulative baseline hazard function. The cumulative baseline hazard function $\Lambda_{0,z}(t)$ is modeled as a finite linear combination of integrated spline basis functions (non-decreasing from 0 to 1) with non-negative coefficients (Wang *and others*, 2015). The advantage of this specification relative to one that is nonparametric is a significant reduction in the dimension of the parameter space while allowing for flexibility. The unknown parameters $\Lambda_{0,z}(t)$ and β_z are estimated by maximizing the likelihood in (3.4) subject to (3.5). Following the method in Wang *and others* (2015), we obtain the maximum likelihood estimates $\widehat{\Lambda}_{0,z}(t)$ and $\widehat{\beta}_z$ by a EM algorithm that involves a two-stage data augmentation with latent Poisson random variables. This method exploits the connection between the proportional hazards model and a non-homogeneous Poisson process.

Plugging $\widehat{\Lambda}_{0,z}(t)$ and $\widehat{\beta}_z$ into Equation (3.5) we obtain an estimator of $S_z(t|W)$ denoted as $\widehat{S}_z(t|W)$. We estimate $F(w|Z=1)$ by its empirical distribution, denoted as $\widehat{F}(w|Z=1)$, based on the covariate information for patients with $Z=1$ in D^*. Since $P[T>t|Z=1]$ can be expressed as

$$P[T>t|Z=1]=\int_w P[T>t|W=w,Z=1]dF(w|Z=1) \tag{3.6}$$

we estimate $P[T>t|Z=1]$ by plugging $\widehat{S}_1(t|W)$ and $\widehat{F}(w|Z=1)$ into Equation (3.6). We denote this latter estimator as $\widehat{P}[T>t|Z=1]$ (Note: $P[T>t|Z=1]$ can be alternatively estimated by the non-parametric Turnbull estimator (Turnbull,

1976) that uses only the outcome information for patients with $Z = 1$.)

Plugging $\widehat{P}[T > t|Z = 1]$ into Equation (3.2) we obtain $\widehat{S}_1(t)$. We obtain $\widehat{S}_0(t)$ from Equation (3.3) by plugging in $\widehat{S}_0(t|W)$ and $\widehat{F}(w|Z = 1)$.

3.2.4 Measure of Treatment Effect

We measure the treatment effect by parsimoniously modeling the relationship between the quantiles of $S_1(\cdot)$ and $S_0(\cdot)$. Specifically, we assume that $S_1^{-1}(p) = \exp(\delta)S_0^{-1}(p)$ for all $0 < p < 1$. This model is equivalent to assuming, for patients with $Z = 1$, an accelerated failure time (AFT) model (Wei, 1992) of the form:

$$\log(T_1) = \log(T_0) + \delta \tag{3.7}$$

Note that $\delta = 0$ implies that $S_1(t) = S_0(t)$ for all t, i.e. the donors have the same distribution of time to event had they not donated.

We estimate δ using the following simulation procedure. Suppose that we are interested in follow up through time τ. We obtain estimates $\widehat{S}_1(t)$ and $\widehat{S}_0(t)$ by the method described in Section 3.2.3. We generate K observations $T_{1,k} \sim \widehat{S}_1(t)$ and another K observations $T_{0,k} \sim \widehat{S}_0(t)$, $k = 1, \ldots, K$. Let $U_{z,k} = \min(T_{z,k}, \tau)$ and $\Delta_{z,k} = I(T_{z,k} < \tau)$ for $z = 0, 1$, $k = 1, \ldots, K$. We then fit model (3.7) using these data. Denote the resulting estimator of δ by $\widehat{\delta}$. We use the R package *aftgee* (Chiou *and others*, 2014) to fit this model and compute $\widehat{\delta}$.

3.2.5 Computation of Standard Errors and Confidence Intervals

We compute estimates of standard error of $\widehat{\delta}$ using nonparametric bootstrap of individuals from the original datasets. In our analysis and simulations, we used 95% Wald-based confidence intervals with the bootstrapped standard error estimator.

3.3 Data Analysis

We apply the methods developed in Section 3.2 to estimate the causal effect of kidney donation on hypertension-free survival and diabetes-free survival among those who chose to donate.

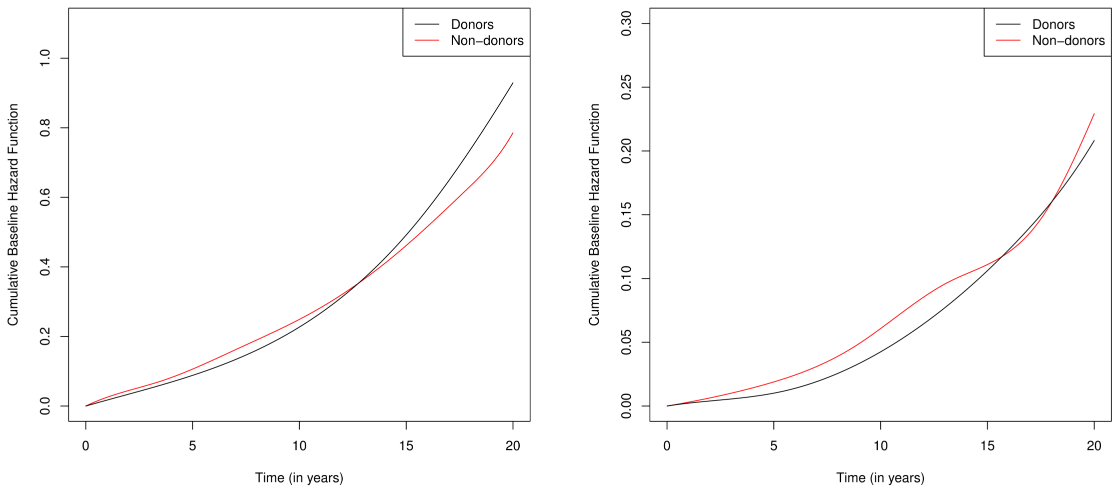
The donors were drawn from the Wellness and Health Outcomes in LivE Donors (WHOLE-DONOR) Study. The earliest of the donations occurred in 1970 and the latest in 2013. Age, gender, race, BMI were measured for each donor at the time of donation. The donors included in the final analytic sample were free of the corresponding disease endpoint at the time of donation. The non-donors were identified from Atherosclerosis Risk in Communities (ARIC) (Visits 1-4; 1987-1998) and Coronary Artery Risk Development in Young Adults (CARDIA) (Visits 1-8; 1985-2011) studies. The non-donors included in the final analytic sample were free of the disease endpoint at the first visit. The last available visit with non-missing disease ascertainment was considered the “end visit”. The preceding visits where the subject is free of the disease endpoint of interest and other co-morbidities (e.g., cardiovascular disease, cancer) were considered valid "enrollments". Age, gender, race and BMI are measured for each subject at each "enrollment". Table 3.1 gives the demographic information of the final analytic samples corresponding to each endpoint. Donors tend to be older, more female, less black and have higher BMI than non-donors.

In the analyses, the cumulative baseline hazard functions were modeled using integrated spline basis functions with five interior knots. The proportional hazards model estimates were not sensitive to selection of the number of interior knots. Further, we set $\tau = 20$ years and $K = 1000$.

3.3.1 Hypertension-free analysis

In the final analytic sample, there are 1,077 live donors and 10,832 eligible non-donors. The non-donors contribute multiple "enrollments" during follow-up. Among the non-donors, 21%, 22%, 45% and 12% contributed 1, 2, 3 and 4 "enrollments" respectively. Among the donors, 12.26% had interval censored observations, 76.42% had right censored observations and 11.32% had exact observations. The percentages of interval censored and right censored observations among the non-donor records were 27.58% and 72.41% respectively.

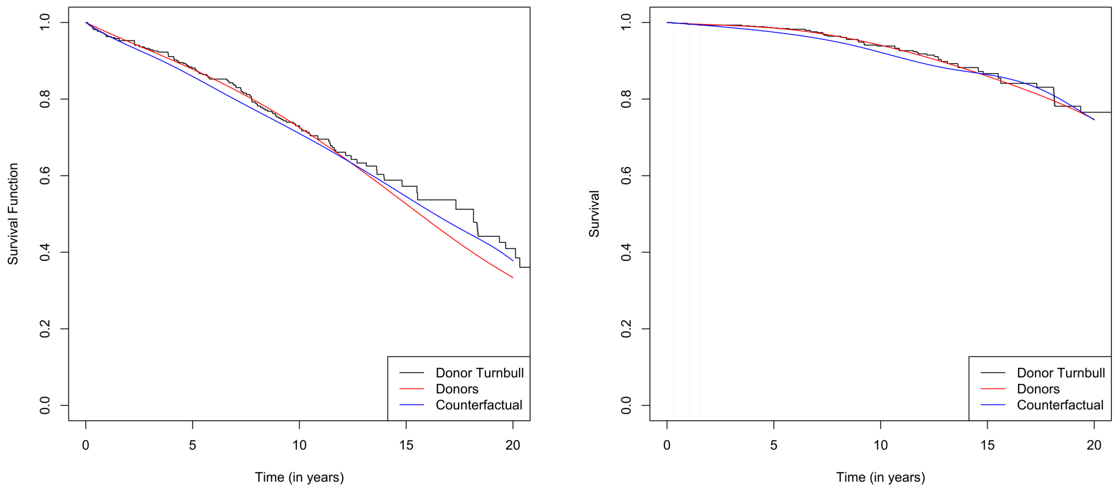
Figure 3.3a shows the estimated cumulative baseline hazard function corresponding to the reference cohort (age 42 years, female, black, BMI 25). Table 3.2 shows the



(a) Endpoint: Hypertension or death

(b) Endpoint: Diabetes or death

Figure 3.3: Estimates of cumulative baseline hazard function in the reference cohort (age 42 years, female, black, BMI 25) for donors and non-donors for the endpoints (a) hypertension or death, and (b) diabetes or death.



(a) Endpoint: Hypertension or death

(b) Endpoint: Diabetes or death

Figure 3.4: Estimates of the donor survival curve, the counterfactual survival curve and the Turnbull estimator among donors for the endpoints (a) hypertension or death, and (b) diabetes or death.

Table 3.1: Demographic information of the subjects in the final analytic sample corresponding to each endpoint.

	Endpoint			
	Hypertension or death		Diabetes or death	
	Donors	Non-donors	Donors	Non-donors
Number of subjects	1077	10,832	1192	9056
Number of “enrollments”	1077	26,597	1192	15,970
Age				
mean	44.32	42.54	44.56	41.17
(sd)	(11.24)	(14.58)	(11.40)	(15.03)
Female (%)	63	55	63	58
Black (%)	13	30	13	31
BMI				
mean	26.43	25.89	26.55	25.10
(sd)	(4.04)	(4.76)	(4.10)	(4.45)

estimated regression coefficients from the PH Model. For donors and non-donors, age, race and BMI were positively and significantly associated with the risk of developing hypertension or dying; gender was not a significant risk factor. Figure 3.4a shows the estimated donor and counterfactual survival curves. For comparative purposes, the Turnbull estimator for donors is also presented. The treatment effect under the AFT model is estimated to be 0.005 [95% CI: -0.10,0.12]. This result may be hard to

Table 3.2: Point estimates and 95% confidence intervals for the regression coefficients obtained from separate PH models for each endpoint.

	Hypertension or death		Diabetes or death	
Covariate	Donors	Non-donors	Donors	Non-donors
Age				
Point estimate	0.054	0.044	0.032	0.043
(95% CI)	(0.041,0.066)	(0.041,0.047)	(0.009,0.056)	(0.035,0.051)
Female				
Point estimate	-0.142	-0.070	-0.006	-0.292
(95% CI)	(-0.392,0.109)	(-0.14,0.0001)	(-0.436,0.425)	(-0.425,-0.158)
Black				
Point estimate	0.509	0.581	0.470	0.687
(95% CI)	(0.194,0.826)	(0.503,0.659)	(-0.164,1.105)	(0.547,0.827)
BMI				
Point estimate	0.067	0.057	0.073	0.067
(95% CI)	(0.037,0.097)	(0.05,0.063)	(0.017,0.128)	(0.055,0.079)

interpret due to the crossing of the estimated survival curves. Table 3.3 reports the estimated differences (and associated 95% confidence intervals) between the donor and counterfactual survival curves at 5, 10, 15 and 20 years. These analyses show that there is no significant evidence to suggest that donors are putting themselves at increased risk for hypertension or death.

3.3.2 Diabetes-free analysis

In the final analytic sample, there are 1,192 live donors 9,056 eligible non-donors. The non-donors contribute multiple "enrollments" during follow-up. Among the non-

Table 3.3: Point estimates and 95% confidence intervals for the difference in the donor and counterfactual survival curves at particular time points for each endpoint.

	Hypertension or death		Diabetes or death	
Time(in years)	Estimate	(95% CI)	Estimate	(95% CI)
5	0.019	(-0.003,0.04)	0.011	(0.004,0.017)
10	0.016	(-0.02,0.053)	0.019	(0.001,0.036)
15	-0.019	(-0.07,0.032)	-0.004	(-0.042,0.033)
20	-0.044	(-0.114,0.027)	0.002	(-0.078,0.081)

donors, 49%, 26% and 25% contributed 1, 2 and 3 “enrollments” respectively. Among the donors, 3.10% had interval censored observations, 92.79% had right censored observations and 4.11% had exact observations. The percentages of interval censored and right censored observations among the non-donor records were 8.32% and 91.68% respectively.

Figure 3.3b shows the estimated cumulative baseline hazard function corresponding to the reference cohort (age 42 years, female, black, BMI 25). Table 3.2 shows the estimated regression coefficients from the PH Model . For non-donors, age, race and BMI were positively and significantly associated with the risk of developing diabetes or dying; gender had a significant negative association. For donors, age and BMI were positively and significantly associated with the risk of developing diabetes or dying; gender and race were not significant risk factors. Figure 3.4b shows the estimated donor survival and counterfactual survival curves; the Turnbull estimator for donors is also presented. The treatment effect under the AFT model is estimated to be -0.004 [95% CI: -0.15,0.14]. Like the hypertension analysis, the estimated survival curves cross. Table 3.3 reports the estimated differences (and associated 95% confidence intervals) between the donor and counterfactual survival curves at 5, 10, 15 and 20 years. At 5 and 10 years, there are statistically significant differences between the donor and counterfactual survival curves, with donation appearing to be protective for the occurrence of diabetes at these time points. At 15 and 20 years, the differences are no longer statistically significant. Donation may be protective for diabetes/death in the early years due possibly to better health care or healthy behavior. Such benefits appear to dissipate in the long term.

3.4 Simulation Results

We conducted a simulation study to evaluate the performance of the proposed methodology. We simulated 500 datasets that closely resembled the data structure for the hypertension-free analysis discussed above. For each dataset, the number of donors was 1,077 and the number of non-donors was 10,832.

3.4.1 Simulation of Donor Data

We generated independent covariates to mimic age, gender (1 = female, 0 = male), race (1 = black, 0 = white) and BMI - age and BMI were simulated as normal random variables with means 44.32 and 26.43 and standard deviations 11.24 and 4.04, respectively, gender and race were simulated as Bernoulli random variables with probabilities 0.63 and 0.13, respectively. Using the 4-dimensional covariate vector W, we generated, for each donor, an exact time-to-event from an exponential regression model with rate $\lambda \exp(\beta^T W)$, $\lambda = 0.01$ and $\beta^T = (0.04, -0.15, 0.42, 0.06)$.

We introduced a censoring mechanism by independently generating four examination times, with the inter-examination times distributed according a truncated exponential distribution with rate 0.25 and truncation at 6 years. When the time-to-event was contained between two examination times, we, with probability 0.8, interval-censored the outcome using the examination times as the end-points and, with probability 0.2, considered the outcome to be exactly observed. If the time-to-event was larger than the time to last examination time, we right-censored the outcome at the last examination time.

3.4.2 Simulation of Non-Donor Data

For non-donors, we generate multiple visit data which will translate into multiple “enrollments”. To start, we generated independent covariates to mimic age (at first visit), gender(1 = female, 0 = male) and race (1 = black, 0 = white). Age was simulated as a normal random variable with mean 42.54 and standard deviation 14.58. Gender and race were simulated as Bernoulli random variables with probabilities 0.55 and 0.30, respectively. For each non-donor, we generated a random variable V, denoting the number of clinic visits (assumed to range from 1 to 4). The probability distribution of V was specified as follows: $P[V = 1] = 0.21, P[V = 2] = 0.22, P[V = 3] = 0.45, P[V = 4] = 0.12$. The duration of time between visits was generated according to a truncated exponential distribution with rate 0.1 and truncation at 20 years. We also generated a $(V + 1)^{th}$ clinic visit using this inter-visit distribution.

Since age is time-varying, we set the age at a given visit to be the age at the first visit plus the time that has elapsed between the given visit and the first visit. We generated BMI at each visit according to a linear mixed effects model with gender, race and visit-specific age as fixed covariates, a fixed intercept, a subject-specific normally-distributed, mean zero random effect and normally-distributed, mean zero random noise. The intercept was set to 17.62, the coefficients for gender, race and age were set to -0.26, 3.04 and 0.17, respectively, and the standard deviations of the random effect and random noise were set to 4.61 and 1.53, respectively.

Our censored outcome data generation process proceeds sequentially by clinic visit. At each visit $v = 1, \ldots, V$ (let t_v be the time of this visit), we generated, using the 4-dimensional covariate vector W (i.e., age, gender, race, BMI) available at this visit, an exact time-to-event from an exponential regression model with rate $\lambda \exp(\delta) \exp(\beta^T W)$, where λ and β^T are the same as specified for the donors and δ is a parameter that differentiates the conditional risk of the event between donors and non-donors. It is important to note that our specification of the exponential regression models for the donors and non-donors implies that (3.7) holds. In our simulation study, we considered $\delta = -0.5, 0, 0.5$. If the exact time-to-event is less than the time to the next visit (let t_{v+1} be the time of this visit), we (1) interval censored the outcome with zero as the left endpoint and the time between visit v and visit $v+1$ (i.e., $t_{v+1} - t_v$) as the right endpoint, (2) stopped the data generation process, and (3) for each previous visit $p = 1, \ldots, v-1$ (let t_p be the time of this visit), we produced an additional enrollment, where the censored outcome has left endpoint $t_v - t_p$ and right endpoint $t_{v+1} - t_p$; otherwise we continued to the next clinic visit. If visit V is reached and the exact time-to-event is not less than the time of visit $V+1$, then we created for each visit $v = 1, \ldots, V$, right censored enrollments with right censoring time $t_{V+1} - t_v$.

3.4.3 Simulation Results

Table 3.4 shows the results of the simulation study, based on 500 simulated datasets. We considered three choices of $\delta = -0.5, 0, 0.5$. For all choices, the bias in estimation of δ is very small and the confidence intervals (constructed using 1000 bootstrap samples) achieve the nominal 95% level. The average of the bootstrapped standard errors of $\widehat{\delta}$ across simulated datasets is approximately equal to the standard deviation of $\widehat{\delta}$'s across these datasets. Overall, the results indicate that the proposed methods perform well in this simulation study.

Table 3.4: Simulation results.

δ	Bias	Standard deviation	Average standard error	Empirical coverage
-0.5	0.008	0.07	0.07	0.96
0	0.005	0.07	0.07	0.95
0.5	0.002	0.07	0.07	0.95

3.5 Discussion

In this paper, we developed an inferential framework for estimating the causal effect among "exposed" subjects on a time-to-event outcome, based on multiple data sources and censored outcome information. This was achieved by conceptualizing and manufacturing a point exposure study that allowed us to identify the causal parameter of interest under certain set of assumptions (i.e., no unmeasured confounders, non-informative censoring).

In our motivating example, the time-to-event outcome was censored in the broadest sense. That is, it was a mix of interval-censored, right-censored and exact observations. With the exception of two working papers (Vandebosch and Goetghebeur, 2005; Valappil *and others*, 2015), we were not able to identify any published causal inference papers with interval-censored outcomes. Our approach relied on specification of a proportional hazards regression model (Cox, 1972). For this model, inference in the presence of interval censoring has been well-studied (see, e.g., Finkelstein, 1986;

Satten, 1996; Goggins *and others*, 1998; Satten *and others*, 1998; Pan, 1999, 2000; Goeteghebeur and Ryan, 2000; Betensky *and others*, 2002; Cai and Betensky, 2003; Zhang *and others*, 2010; Wang *and others*, 2015 for frequentist approaches and Sinha *and others*, 1999; Yavuz and Lambert, 2011; Wang *and others*, 2013; Lin *and others*, 2015 for Bayesian approaches). In our setting (large number of subjects, some with exact observations), the majority of these approaches are too computationally expensive or too technically complicated to practically implement. Before adapting the approach of Wang *and others* (2015), we experimented with the `R` packages `intcox` and `coxinterval`. The package `intcox` adopts the iterative convex minorant approach of Pan (1999) but it also produces biased parameter estimates as pointed out by Wang *and others* (2015). The package `coxinterval` developed based on Boruvka and Cook (2015) could not be easily adapted to handle exact observation times.

Our analysis can be prone to bias if the underlying assumptions (i.e., no unmeasured confounding, non-informative censoring) are violated. In terms of assumptions, we are most concerned about the no unmeasured confounding assumption. Our ability to adjust for measured confounding factors is limited by the fact that we require that all data sources record data on the same set of factors. In our analysis, we adjusted for gender, race, age and BMI which were well recorded in live donor database and the ARIC and CARDIA studies. However, it is likely that there are additional confounding factors at play, e.g., blood pressure and glomerular filtration rate. While these factors were to be recorded in the multiple datasets, they have missing data rates of the order of 15%. In future work, we plan to extend our methods to handle this issue.

Another limitation of our methodology is model specification. Specifically, our analysis relies on correct specification of proportional hazards regression models for the time-to-event for donors and non-donors. This may lead to some bias in the estimate of the target causal parameter. In a future work, we plan to explore methods that are more robust to such misspecification. Our proposed estimator of the treatment effect relies on an accelerated failure time model that connects the donor and counterfactual survival times. This assumption may not hold, especially in settings whether the associated survival curves cross. Nonetheless, the estimator can be thought of as the best fitting accelerated failure time model that is consistent with the estimated survival curves. Future work will explore alternative methods for contrasting the survival curves.

In evaluating exposure effects, it is not uncommon for information on exposed and non-exposed subjects to be obtained from different data sources. The methods developed in this paper should be useful for evaluating such effects, provided that (1) one can conceptualize a hypothetical point exposure study and (2) the underlying data sources collect a common set of confounding factors.

Chapter 4

Testing Equality of Curves After Covariate Adjustment

4.1 Introduction

We propose simple methodological approaches for global and local tests of the difference between the mean of treatment and control groups when the measured outcome is a function. Several papers in the functional data analysis literature have focussed on comparing the averages of two functional processes. For example, Benko *and others* (2009) developed bootstrap-based tests of equality of means, eigenvalues and eigenfunctions of the covariance function in the two sample problem. Hall and Keilegom (2007) used bootstrap-based tests for equality of distributions of two independent samples of curves. Zhang *and others* (2010) proposed L^2-based and bootstrap-based statistics for testing equality of two average curves when the subject specific curves are independent and observed without noise. Crainiceanu *and others* (2012) proposed a bootstrap-based inference procedure for the difference in means of two correlated functional processes. However, none of these approaches considered covariate-adjusted testing, which is essential in cases when covariates may differ across groups. Several authors have developed Bayesian approaches for this problem in settings with complex correlation structures (Behseta and Kass, 2005; Behseta *and others*, 2007; Morris *and others*, 2003; Morris and Carroll, 2006; Morris *and others*, 2006, 2011).

The scientific problem that motivated our study is whether targeted deletion of interleukin 10 gene (IL-10^{tm1Cgn}) in mice leads to decrease in oxygen consumption of the animal. We have repeated measures of oxygen consumption in a group of 10 mice where the gene has been knocked out and a control group of 10 mice where the gene is present. The measurements were taken at regularly spaced time points over four days. We want to explore if the average oxygen consumption through the day (midnight-midnight) differ significantly between the groups and if the genotype-outcome association is altered by the body composition of the animal. The novelty of our approach is that it addresses the problem that each animal has repeated functional measurements over multiple days (oxygen consumption measure at every 30 minutes for 4 days) and additional covariates of interest (i.e., body composition measures).

We develop a permutation based approach to test for a global difference between the averages of two functional processes after covariate adjustment using the estimated L^2 area under the squared difference curve as the test statistic. We also test for localized differences between the two covariate adjusted average curves using the 95% pointwise and joint confidence intervals obtained using a nonparametric bootstrap of subjects. The main novelty of our paper is that we are using the covariate adjusted curves to develop the test procedures and take into account the within-subject sampling functional correlation. The proposed approach is easy-to-implement, computationally fast and scalable and adaptable to more complex settings. In Section 4.2 we develop the statistical framework for our method. Section 4.3 provides the results of the real data analysis and the simulation study. We conclude with a discussion in Section 4.4.

4.2 Methods

Our method utilizes information from two data sources having similar structure: the first one comes from "treated" animals (e.g., IL10^{tm} group) and the second one comes from animals who are "not treated" (e.g., control group). Both data sources have information on a functional outcome [e.g., oxygen consumption measured at regular intervals (∼30 minutes) over a period of time (4 days)] and baseline covariates (e.g., body composition measures). Figure 4.1 displays the scatter plots for oxygen consumption of the animals in each group in a 24 hour period (midnight-midnight)

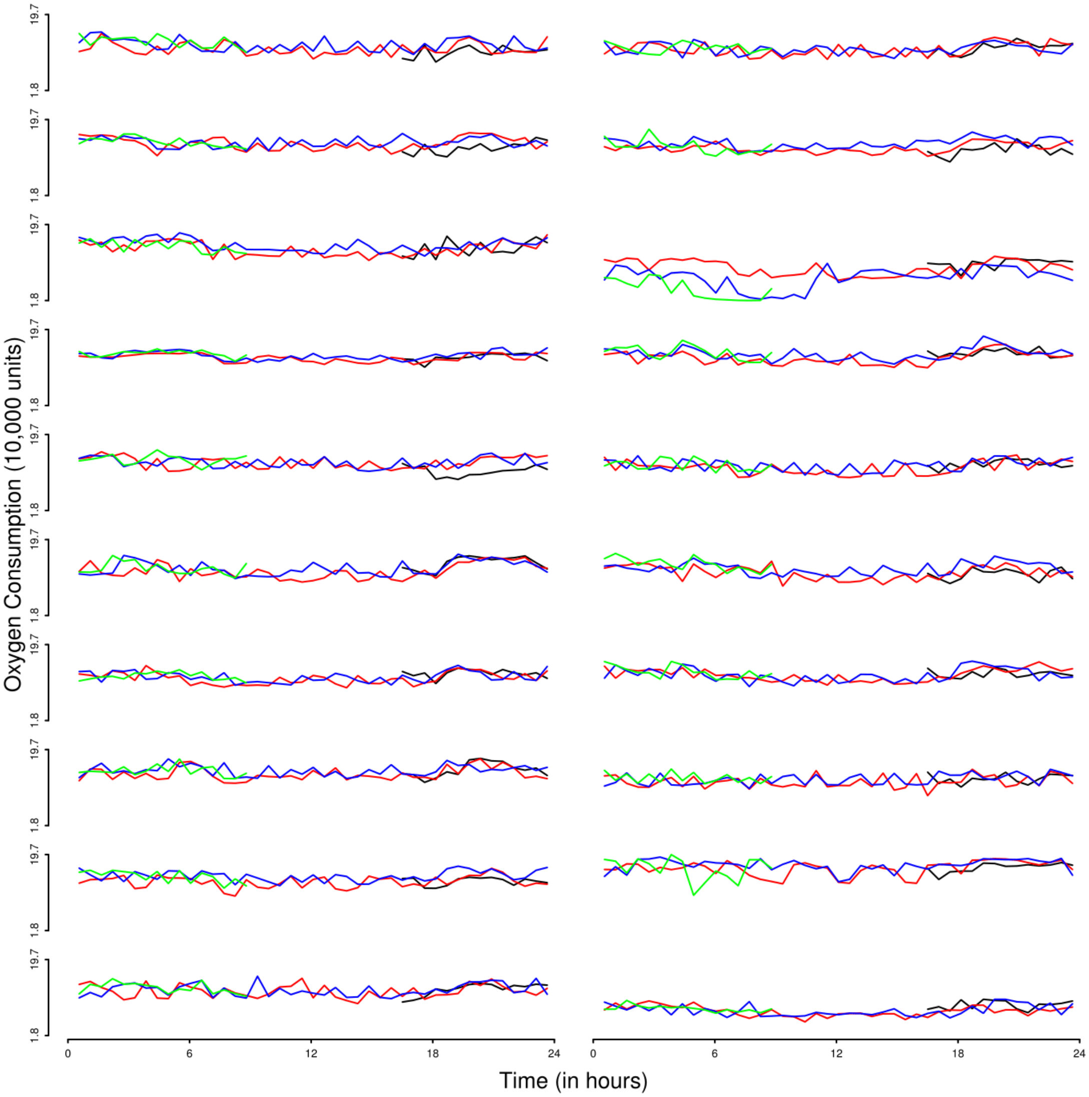


Figure 4.1: Plot of oxygen consumption during the 24 hours over multiple days; the panels in the left correspond to animals in the control group and panels on the right correspond to animals in the IL-10^{tm1Cgn} group. Each panel shows the oxygen consumption of an animal over 4 days: Day 1 (black line), Day 2 (red line), Day 3 (blue line) and Day 4 (green line).

over 4 days. Figure 4.2 displays the average oxygen consumption (over every observation within days and all four days) as a function of body mass composition as well as the body mass composition distribution within treatment group. Let n_i denote the number of animals in the i^{th} group, $i = 0$ for "treated" animals and $i = 1$ for "not treated" animals. We observe $\{(Y_{ijl}(t), X_{ij}) : t = t_1, \ldots, t_k; i = 0, 1; j = 1, \ldots, n_i; l = 1, 2, 3, 4\}$, where $Y_{ijl}(t)$ denotes the functional outcome observed at the time points $t_1, \ldots, t_k$ in the range $[0, T]$ during the l^{th} day and X_{ij} denotes the vector of baseline covariates for the j^{th} animal in the i^{th} group. Denote by $Y_{ij.}(t) = \dfrac{1}{4}\sum_{l=1}^{4} Y_{ijl}(t)$. We are interested in a model of the type:

$$Y_{ij.}(t) = \beta_{i0}(t) + X_{ij}\beta_{i1}(t) + \epsilon_{ij}(t) \tag{4.1}$$

where $\epsilon_{ij}(t)$ is a mean zero process with unspecified correlation structure. We want to test the hypothesis: $\mu_1(t) = \mu_0(t)$, where $\mu_i(t) = E[Y_{ij.}(t)]$ for $i = 0, 1$. Note that $\mu_i(t) = E[Y_{ij.}(t)] = E[E[Y_{ij.}(t)|X_{ij}]] = \beta_{i0}(t) + \beta_{i1}(t)E[X_{ij}]$. We model the functional regression parameters in equation 4.1 using P-splines that combine a B-spline basis with a discrete penalty on the basis coefficients (Eilers and Marks, 1996). The regression functions are estimated using restricted maximum likelihood estimation of the associated penalized least squares objective function in the framework of generalized additive models (Chambers and Hastie, 1991; Hastie and Tibshirani, 1990). For $i = 0, 1$ we estimate $\mu_i(t)$ by $\widehat{\mu}_i(t) = \widehat{\beta}_{i0}(t) + \widehat{\beta}_{i1}(t)\widehat{E}[X_{ij}]$, where $\widehat{E}[X_{ij}]$ is the sample average of the covariates in the i^{th} group. We are making the working assumption that $\epsilon_{ij}(t)$ are independent. This assumption substantially simplifies the estimation procedure, though for inference we will take the within-subject correlation into account. Estimating parameters under independence and then correcting the confidence intervals has a long and successful history in statistics.

4.2.1 Test for Global Difference

Define $\delta(t) = \mu_1(t) - \mu_0(t)$ and denote by $I = \int_0^T \delta^2(t)dt$. Note that $\delta(t) = 0$ for every t if and only if $I = 0$. Thus I is a measure of global difference between the means of the two groups. Denote the estimates of the within group averages by $\widehat{\mu}_i(t), i = 0, 1$ and let $\widehat{\delta}(t) = \widehat{\mu}_1(t) - \widehat{\mu}_0(t)$ be an estimator of $\delta(t)$. We estimate I by the Riemann sum approximation: $\widehat{I} = \frac{T}{K+1}\sum_{i=0}^{K}\widehat{\delta}^2(w_i)$, where $w_0 = 0, w_1, \ldots, w_K = T$ is a fine grid of equally spaced points on $[0, T]$ and K is a large number. We consider the null hypothesis: $H_0 : I = 0$ versus the alternative hypothesis $H_a : I > 0$. Our permutation based test procedure involves the following steps:

(i) Consider the joint dataset with $n = n_1 + n_0$ animals, where for $i = 0, 1$, n_i animals come from i^{th} group. Consider a random permutation p of the labels of "treatment" (i.e. "treated" or "not treated").

(ii) For the permuted dataset, estimate the averages of the two groups: $\mu_1^{(p)}(t)$, $\mu_0^{(p)}(t)$ by the model fitting and estimation procedure described earlier. Denote the difference function $\delta^{(p)}(t) = \mu_1^{(p)}(t) - \mu_0^{(p)}(t)$ and the integral of the squared difference function by I_p. Compute $\widehat{I_p}$ by the method described earlier.

(iii) Repeat step (i) with P permuted datasets.

(iv) Compute the permutation test p-value to be the proportion of permutations with $\widehat{I_p} \geq \widehat{I}$.

The key idea of the permutation test is as follows: under the null hypothesis there is no difference in the average outcome between the two groups. Hence the treatment labels are exchangeable under the null hypothesis. The empirical distribution of $\widehat{I_p}$ estimates the distribution of the global difference between the two groups under the null hypothesis. This provides the rationale for the computation of the test p-value in step (iv).

4.2.2 Test for Localized Differences

We also propose a test for localized differences between groups, (i.e., the difference in average outcome at particular time points) using a nonparametric bootstrap-based inferential procedure (Crainiceanu *and others*, 2012). The main difference from the procedure in Crainiceanu *and others* (2012) is that we are working with covariate adjusted curves, which is important in many applications.

One question of interest is whether there is a difference in the average outcomes between the groups at a fixed time point t. The corresponding null and alternative hypotheses can be stated as:

$$H_{0,t} : \mu_1(t) = \mu_0(t) \ \text{ versus } \ H_{a,t} : \mu_1(t) \neq \mu_0(t) \ \text{ for a fixed } t \tag{4.2}$$

We compute the 95% pointwise confidence intervals to address this question.

Another question of interest is whether there is a difference between the average curves at all time points. The corresponding null and alternative hypotheses are as follows:

$$H_{0,m} : \mu_1(t) = \mu_0(t) \ \ \forall t \text{ versus } \ H_{a,m} : \mu_1(t) \neq \mu_0(t) \ \text{ for at least one } t \tag{4.3}$$

This question can be addressed using the 95% joint confidence intervals to account for multiple hypotheses testing.

The key steps in the computation of the different kind of confidence intervals are as follows:

(i) Generate B simple random samples with replacement separately from each group.

(ii) For each bootstrap dataset, define $\delta_b(t) = \mu_{1b}(t) - \mu_{0b}(t)$, $b = 1, \ldots, B$. Estimate $\mu_{1b}(t)$, $\mu_{0b}(t)$, and $\delta_b(t)$ by the procedure described earlier.

(iii) Compute 95% pointwise and joint confidence intervals.

The 95% pointwise confidence intervals in step (iii) are constructed based on the bootstrap distribution of $\delta_b(t)$ for a fixed t. More specifically, we estimate the standard error of the difference of means based on the bootstrap samples and use the z-score cutoff. They can be interpreted as follows: at each time point t in repeated samples the true difference will be covered by the interval 95% of the time. The 95% joint confidence intervals are computed by the algorithm given in Section 3 of Crainiceanu *and others* (2012). The interpretation is as follows: at all time points in repeated samples the true difference will be covered by the interval 95% of the time.

4.3 Results

A mouse with targeted deletion in the interleukin 10 gene (IL-10^{tm1Cgn}) has been proposed as a mouse model for frailty and low-grade inflammation. The older frail IL-10^{tm} mice show many similarities with older frail human beings. This provides the rationale for using it as a scientific model for studying frailty. It has been hypothesized that older, frail mice have decreased oxygen consumption compared to the normal wildtype mice. We use the methods developed in the Section 4.2 to explore the validity of this hypothesis and to investigate whether the statistical association between the genotype and decreased oxygen consumption is altered by the body composition of the animal.

4.3.1 Description of study design and data

We have experimental data on $n_0 = 10$ mice with the interleukin 10 gene knocked out (IL-10^{tm} group) and $n_1 = 10$ additional mice where the gene is present (control group). For the animals in each group we have repeated measures of oxygen consumption per gram body weight (every 30 mins over 4 consecutive days; 116 repeated observations for each animal, cf Figure 4.1). We also have information on body composition measures (body weight, lean mass, fat mass, fluid mass) for each animal obtained through Nuclear Magnetic Resonance (NMR) experiments. We want to compare the average daily oxygen consumption curves between the groups after adjusting for body composition measures.

4.3.2 Exploratory Analysis, Outlier Identification and Covariate Adjustment

Figure 4.2 displays the bivariate distributions of oxygen consumption and the lean mass and fat mass in the two genotype groups. The upper panel includes observations at all time points as the dependent variable. The lower panel uses the average oxygen consumption over time as the dependent variable. For illustrative purposes we used thin plate regression spline smoothing with four basis functions. The oxygen consumption vs fat mass relationship is similar in two genotype groups. However, there is a difference in the oxygen consumption vs lean mass relationship between the two genotype groups.

Figure 4.3 displays the relationship between lean mass and fat mass for both groups of animals using a thin plate regression spline smoother with four basis functions. We identified one outlying animal in the IL-10^{tm} group with lean mass 22.4 grams and fat mass 1.7 grams. The striking difference in the nature of the red curve in the upper and lower panel of Figure 4.3 supports this fact. In addition to the main analysis (Analysis I) that includes all the animals, we also perform a sensitivity analysis excluding this outlying animal (Analysis II).

Figure 4.4 displays the bivariate distribution of oxygen consumption and the ratio of fat mass and lean mass in the two genotype groups for both Analyses I and II. The upper panel includes observations at all time points as the dependent variable. The lower panel uses the average oxygen consumption over time as the dependent variable. We use thin plate regression splines with four basis functions to smooth the data. The animals in the IL-10^{tm} group have lower values of the ratio of fat mass and lean mass compared to the animals in the control group. These plots also indicate that one animal may be an outlier.

The exploratory analyses indicate that the ratio of fat mass and lean mass is a key body composition measure that could potentially mediate the association between genotype and oxygen consumption. For the rest of the analysis, we use the empirical average of the oxygen consumption at a particular time over the four days as our functional outcome and the ratio of fat mass and lean mass as the scalar covariate

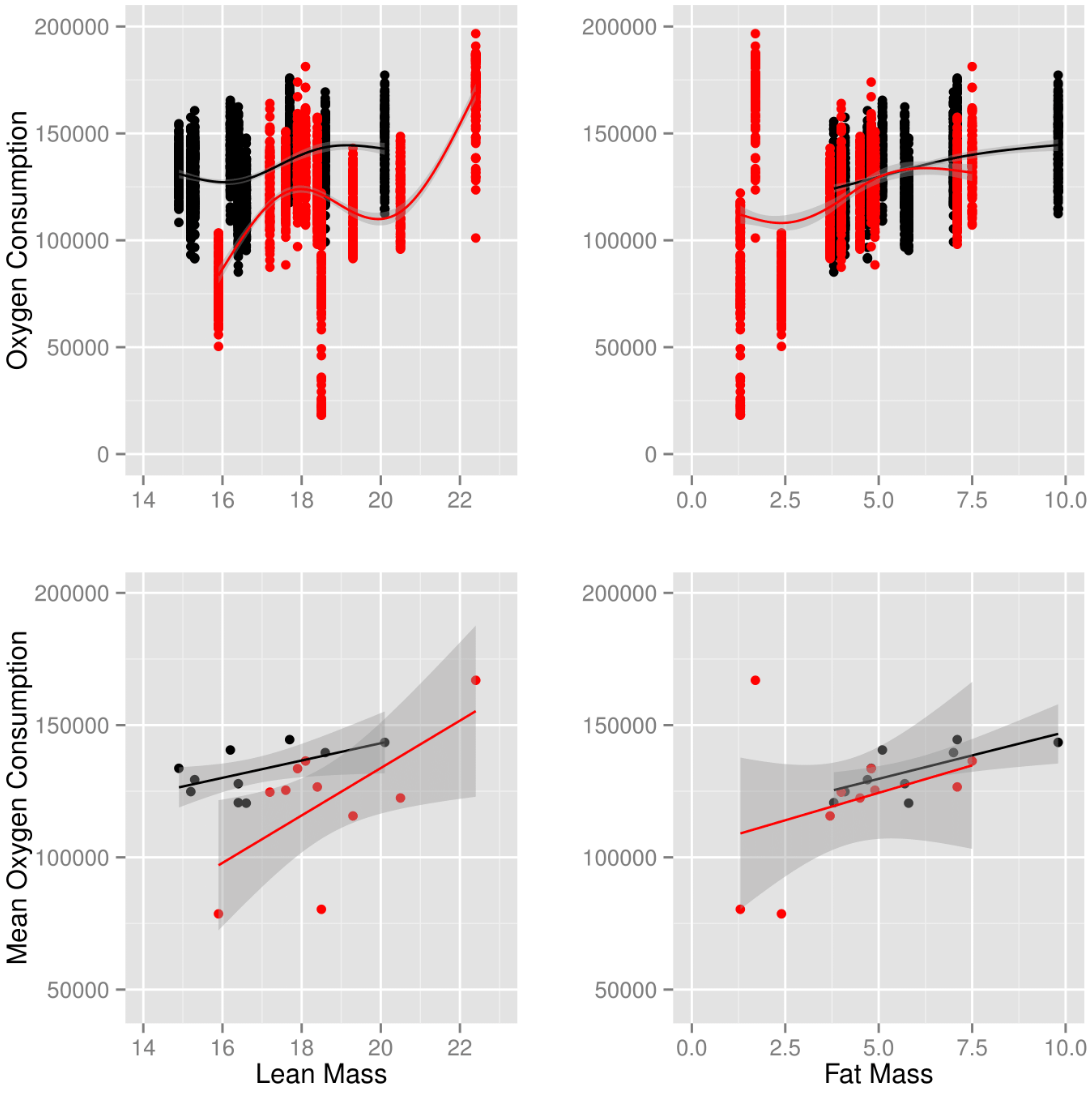


Figure 4.2: Bivariate relationship of oxygen consumption with lean mass and fat mass in the two genotype groups (black color for control group and red color for IL-10^{tm} group): in the upper panel the dependent variable is observed oxygen consumption at all time points; in the lower panel the dependent variable is average oxygen consumption over time.

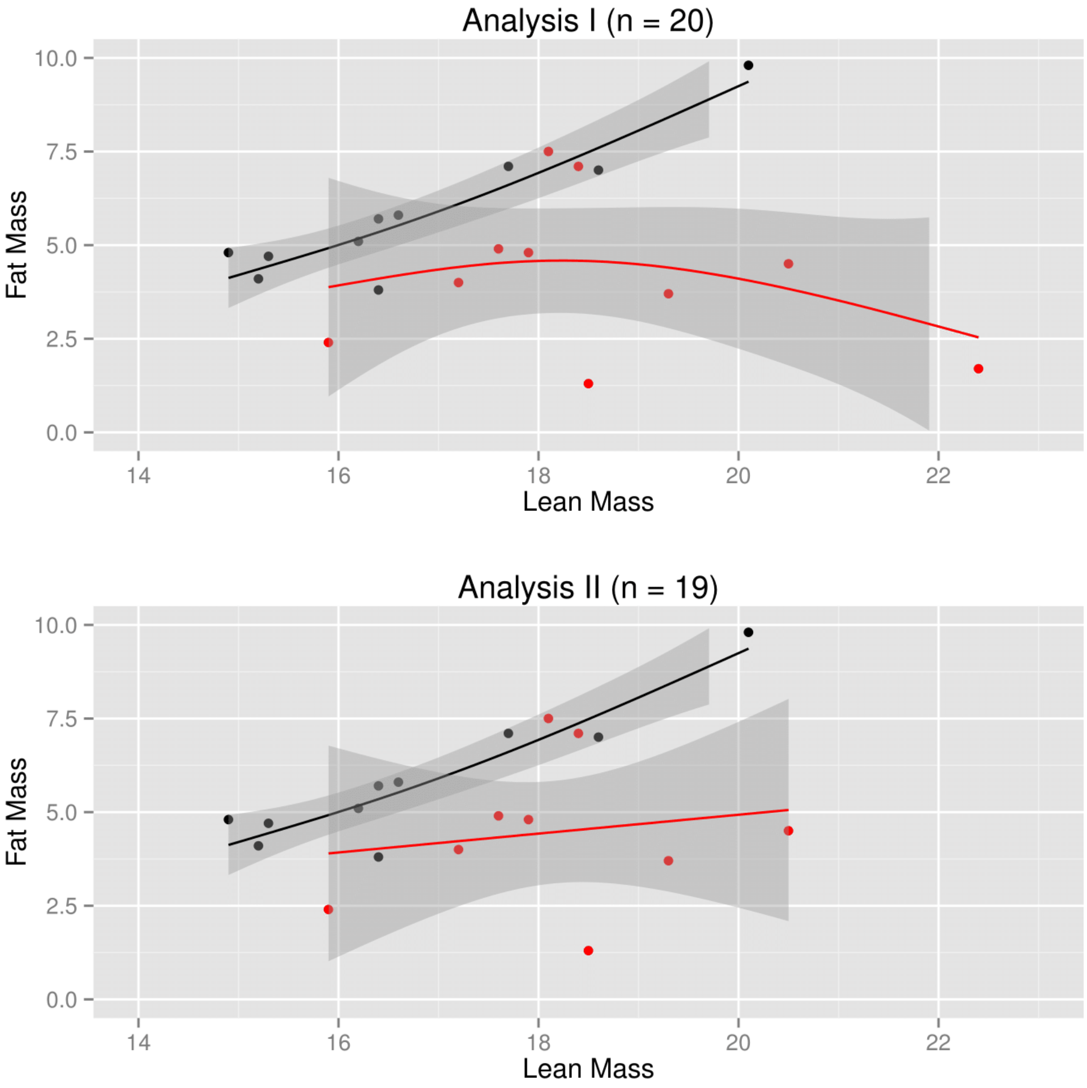


Figure 4.3: Relationship between lean mass and fat mass in the two genotype groups (black color for control group and red color for IL-10^{tm} group): upper panel corresponds to Analysis I ($n = 20, n_0 = 10, n_1 = 10$) with the outlier, lower panel corresponds to Analysis II ($n = 19, n_0 = 9, n_1 = 10$) without the outlier.

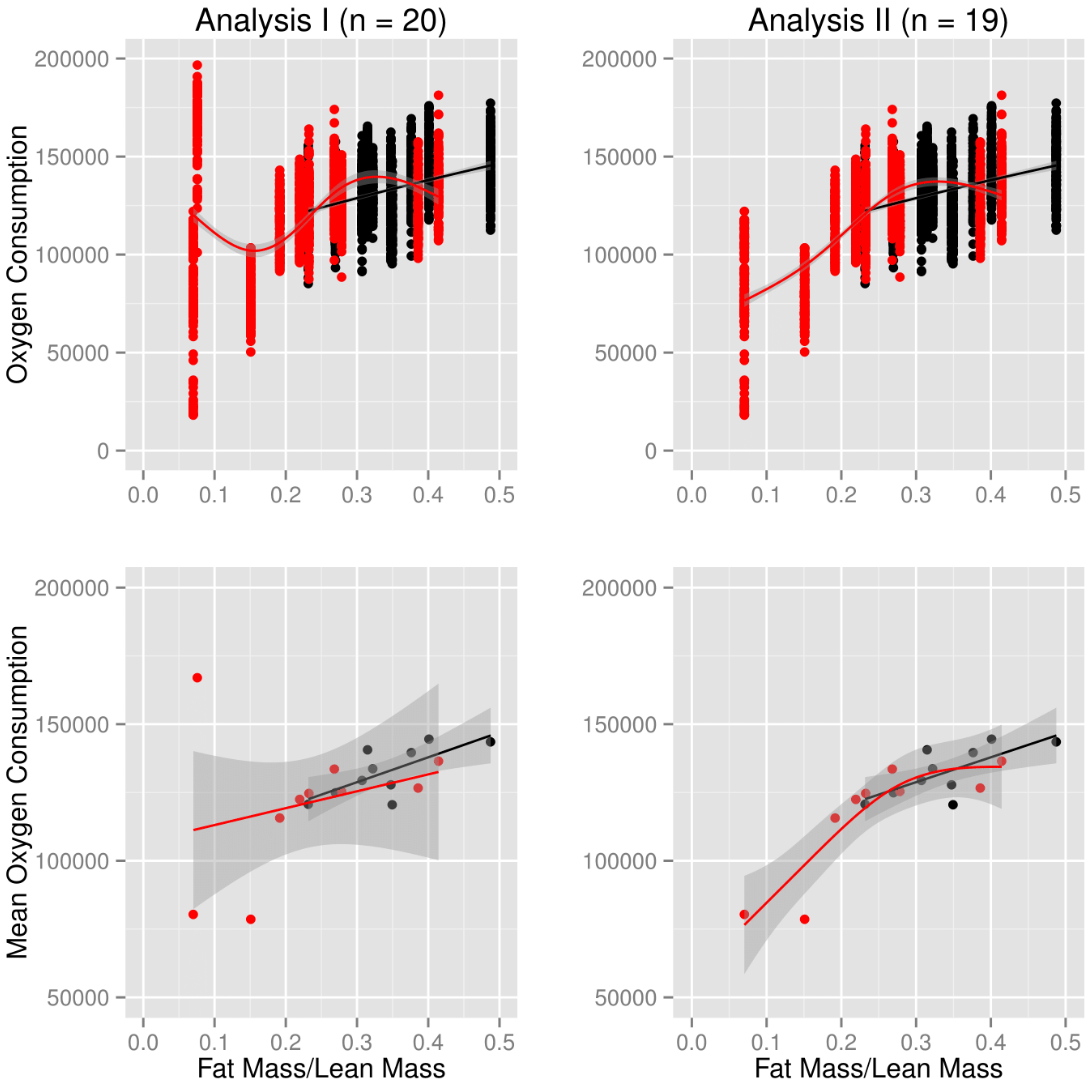


Figure 4.4: Bivariate relationship of oxygen consumption with ratio of fat mass and lean mass in the two genotype groups (black color for control group and red color for IL-10^{tm} group) for both Analysis I ($n = 20, n_0 = 10, n_1 = 10$) and Analysis II ($n = 19, n_0 = 9, n_1 = 10$): in the upper panel the dependent variable is observed oxygen consumption at all time points; in the lower panel the dependent variable is average oxygen consumption over time.

of interest. The covariate adjusted curves are computed by the methods described in Section 4.2. To investigate the potential mediation hypothesis, we compare results with the approach that normalizes the outcome by per gram body weight. The latter approach is routinely applied in studies of mouse metabolism (Speakman, 2013).

4.3.3 Global Genotype Effect

We follow the permutation based approach outlined in Section 4.2.1 to test for global genotype effect. We estimate the global difference over a fine grid on the range [0,24 hours] (i.e., midnight-midnight) where the time points are 0.01 hours apart. We compute the p-value based on 1000 permutations. Note that when we use the body weight normalized outcome the permutation test results provide evidence for significant global genotype effect ($p\text{-}value = 0.01$). However, after adjusting the outcome with the ratio between fat mass and lean mass, there is no such evidence ($p\text{-}value = 0.19$).

4.3.4 Localized Genotype Effect

Figure 4.5 displays the point estimates and the 95% pointwise and joint confidence intervals for the difference in average oxygen consumption between the control mice and the IL-10^{tm} mice. The results are based on $B = 1000$ bootstrap samples. When the oxygen consumption is normalized per gram of body weight (left panel) we observe a significant decrease in average oxygen consumption at different times during the day for the IL-10^{tm} mice compared to the control mice. However, when oxygen consumption is adjusted for the ratio of fat mass and lean mass of the animal, the difference is no longer statistically significant.

Both Analyses I and II resulted in similar findings, indicating that results are not strongly influenced by the one outlier identified in the exploratory process. The results in Sections 4.3.3 and 4.3.4 provide evidence that supports the hypothesis that the association of interleukin 10 gene deletion on the average oxygen consumption is mediated by the ratio of fat mass and lean mass of the animal. The total computation time for performing the tests for global genotype effect and localized genotype effect was around 10 mins (Quad Core Processor 2.2 GHz, 8 GB RAM Macbook Pro)

4.3.5 Simulation Results

We investigate the performance of the proposed methods in a simulation study. For different settings we generate 500 datasets from Model 4.1 with a single covariate, for different choices of $\beta_{i0}(t)$ and $\beta_{i1}(t)$. We consider a time grid of 100 equally spaced points in the interval [0,1]. We generate $\epsilon_{ij}(t)$ in Model 4.1 from a Gaussian Process distribution characterized by the equation $\epsilon(t) = \sum_{k=1}^{4} \xi_k \phi_k(t)$ where ξ_k are mutually independent $N(0, \lambda_k)$ for $k = 1, 2, 3, 4$ and λ_k and $\phi_k(t)$ represent the k^{th} eigenvalue and eigenfunction respectively of the functional principal component decomposition of a centered and scaled version of outcome data. We consider different settings that

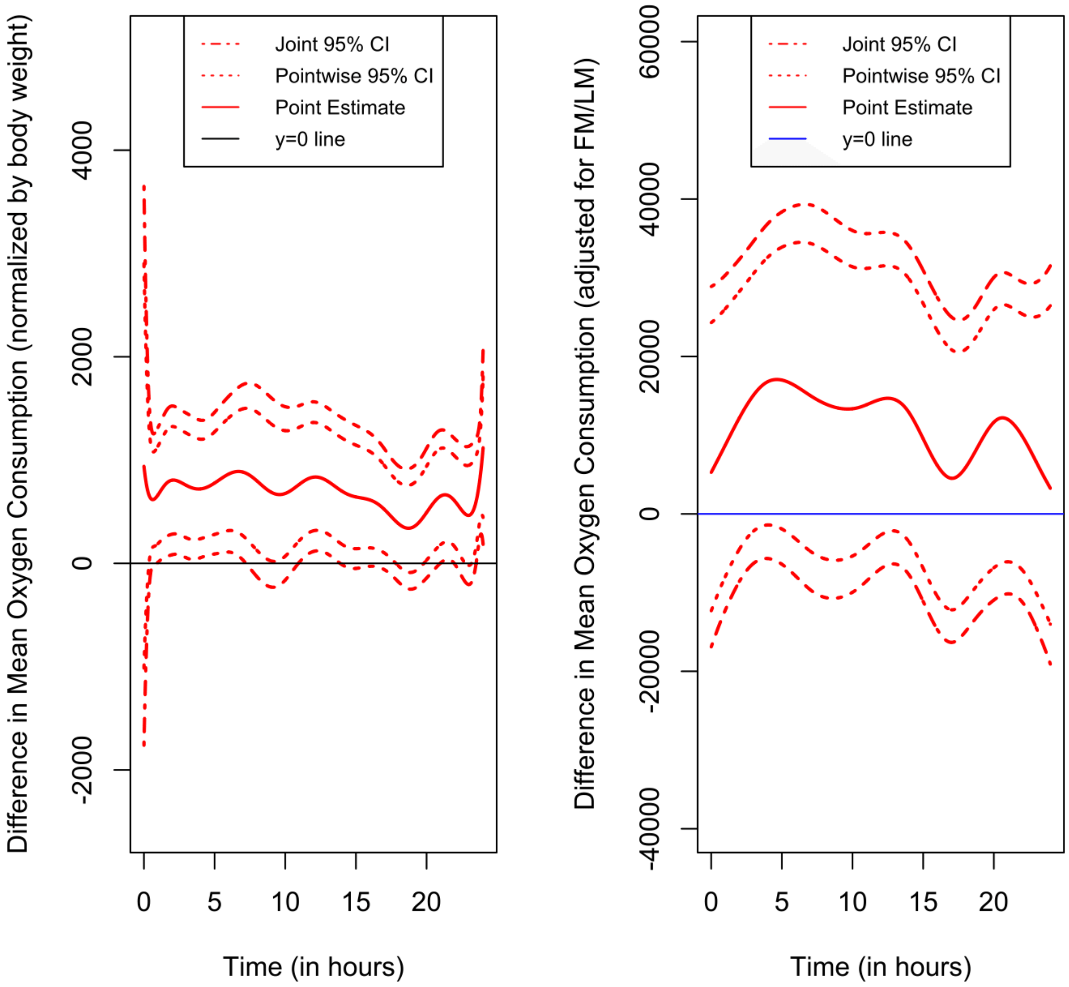


Figure 4.5: Plots showing 95% pointwise and joint confidence intervals for the difference in mean oxygen consumption between control mice and IL-10^{tm} mice: in the left panel the oxygen consumption is normalized by body weight (BW) and in the right panel the oxygen consumption is adjusted for the ratio of fat mass and lean mass (FM/LM) of the animal.

Table 4.1: Simulation results: Scenario 1 corresponds to $\beta_{00}(t) = \beta_{10}(t) = 5$ and $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and covariate distribution in both the groups same as in the control group of the motivating example; Scenario 2 corresponds to $\beta_{00}(t) = \beta_{10}(t) = \sin \pi t$, $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and covariate distribution in both the groups same as in the control group of the motivating example; Scenario 3 corresponds to $\beta_{00}(t) = 5, \beta_{10}(t) = 5.4$ and $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and covariate distribution in both the groups same as in the control group of the motivating example; Scenario 4 corresponds to $\beta_{00}(t) = 0.1(1+t)^2, \beta_{10}(t) = 0.5(1-t)^2$ and $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and covariate distribution in both the groups same as in the control group of the motivating example.

			Global Test	Local Test	
1 - α	Scenario	n	$\widehat{P}$	$\widehat{IAC}_P$	$\widehat{IAC}_J$
0.95	1	10	0.05	0.89	0.87
		20	0.04	0.92	0.92
		50	0.05	0.94	0.94
		100	0.05	0.95	0.95
	2	10	0.05	0.88	0.85
		20	0.04	0.92	0.9
		50	0.05	0.94	0.93
		100	0.05	0.95	0.95
	3	10	0.79	0.89	0.87
		20	0.99	0.92	0.92
		50	1	0.94	0.94
		100	1	0.95	0.95
	4	10	0.65	0.89	0.87
		20	0.99	0.92	0.91
		50	1	0.94	0.94
		100	1	0.95	0.95

combine the choices of the following parameters:

1. Number of subjects: consider $n_1 = n_0$ with $n = n_1 + n_0$ and take $n = 10, 20, 50, 100$.

2. For the regression functions consider two scenarios for data generated under the null hypothesis i.e., $\delta(t) = 0$: (i) $\beta_{00}(t) = \beta_{10}(t) = 5$ and $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and generate covariate from the same distribution for each group (e.g., we use empirical distribution of the covariate values in the control group); (ii) $\beta_{00}(t) = \beta_{10}(t) = \sin \pi t$, $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and generate covariate from the same distribution for each group (e.g., we use empirical distribution of the covariate values in the control group); two additional scenarios for data generated under the alternative hypothesis i.e., $\delta(t) \neq 0$: (iii) $\beta_{00}(t) = 5, \beta_{10}(t) = 5.4$, $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and generate covariate from the same distribution for each group (e.g., we use empirical distribution of the covariate values in the control group); (iv) $\beta_{00}(t) = 0.1(1+t)^2, \beta_{10}(t) = 0.5(1-t)^2$, $\beta_{01}(t) = \beta_{11}(t) = 0.2$ and generate covariate from the same distribution for each group (e.g., we use empirical distribution of the covariate values in the control group).

For the test of global differences, let P be the conditional probability that the null hypothesis is rejected given the true data generating mechanism. When data are generated under the null hypothesis, P is the probability of type I error. When data are generated under an alternative hypothesis, P is the power of the test under that particular alternative. We estimate P by $\widehat{P}$, the proportion of the simulated datasets for which the test procedure rejects the null hypothesis (i.e., permutation test p-value < 0.05). For the test of localized differences we estimate the integrated actual coverage for pointwise confidence intervals (IAC_P) and the integrated actual coverage for joint confidence intervals (IAC_J) as described in Crainiceanu *and others* (2012).

The global test produces the right Type I error for a sample size as little as $n = 10$ as shown in the scenarios 1 and 2 in Table 4.1 where data are generated under $\delta(t) = 0$. In scenarios 3, 4 data are generated under $\delta(t) \neq 0$. For scenario 3, the global test rejects the global null hypothesis 79% of the 500 tests when the sample size is $n = 10$. This improves with increasing sample size; for $n = 20$ the global null hypothesis is rejected 99% of the tests and for larger sample sizes (i.e., n = 50, 100) it is rejected in all cases. For scenario 4, the global null hypothesis is rejected in only 65% of the cases for $n = 10$. For higher sample sizes ($n = 20, 50, 100$) the characteristics are similar to Scenario 3. The 95% pointwise and joint confidence intervals suffer from under-coverage for the case $n = 10$ but coverages improve with increasing sample size.

4.4 Discussion

In this paper, we provide simple and fast methods for testing if and where two covariate adjusted average curves are different.

One question we wanted to explore was whether there is an overall difference between the covariate adjusted curves. We develop a simple easy-to-implement and novel test procedure by adapting the permutation test idea to functional outcomes. To the best of our knowledge, this is the first time such an approach is being proposed in the context of testing equality of two curves after covariate adjustment. We also propose a test for localized difference between the genotype groups (i.e., difference in average outcome at particular time points) using a non-parametric bootstrap of subjects (Crainiceanu *and others*, 2012). The methods in Crainiceanu *and others* (2012) were developed for a matched case-control study. The major difference between our approach and the one presented in Crainiceanu *and others* (2012) is that we are working with covariate adjusted curves.

The issue of covariate adjustment is of great importance in most of the scientific problems for reducing the variability and adjusting for baseline imbalances. For instance, Figure 4.4 shows that the two genotype groups differ significantly with respect to the ratio of fat mass and lean mass. One of the strengths of our approach is that the we perform the covariate adjustment in a statistically principled way: we first model the relationship of the oxygen consumption function and the ratio of fat mass and lean mass using restricted maximum likelihood based functional regression methods; and then use the estimated regression functions and the within group sample averages of the covariate to estimate the within group average oxygen consumption. One issue of concern is that the empirical distribution of the ratio of fat mass and lean mass has limited overlap between the genotype groups.

The global test shows good performance in the simulation studies in terms of Type I error and power. However, the 95% point-wise and joint confidence intervals suffer from under-coverage for low sample sizes (e.g., $n = 10$); but the coverage improves with increase in sample size. For the 95% point-wise confidence intervals, if we use the cutoff based on t distribution with $(n/2 - 1)$ degrees of freedom as opposed to the regular z-score cutoff, we get substantial improvement in coverage for the case $n = 10$ and $n = 20$ under all scenarios. For higher sample sizes (i.e. $n = 50, 100$) the coverages with t distribution cutoff and $z-$score cutoff assume similar values and this finding is also consistent across all scenarios. The algorithm in Crainiceanu *and others* (2012) for producing joint confidence intervals assumes multivariate normality of the difference function. We tried other options, e.g., using a multivariate t distribution with $(n/2 - 1)$ df or the empirical distribution based on bootstrap as suggested by Crainiceanu *and others* (2012). However, both these approaches result in marginal improvement in coverage of the joint confidence intervals for low sample sizes (e.g., $n = 10$). In a future work, we plan to develop methods to handle this issue.

We also explored the sensitivity of the developed methods to outliers. The exploratory analyses have identified one outlying animal in the IL-10^{tm1Cgn} group. We performed a sensitivity analysis by excluding this animal. However, the results were very similar to the main analysis. Thus outlying animal does not strongly impact the findings.

In summary, the key advantages of this method are its ease of implementation, efficiency and scalability. Although it is targeted to address the scientific question posed by the specific application, it can be adapted to a wide variety of biomedical and public health settings with similar design and data structure.

Chapter 5

Conclusion

In this dissertation, we followed the scientific discovery process (Langley, 1987). First, we identified, via collaborations, important scientific questions and the data sources available. The questions included:

1. Do asthma and obesity have a common genetic risk factor (i.e., *ORMDL3* locus on Chromosome 17)? (Chapter 2)

2. Are kidney donors at risk for adverse health consequences (e.g., hypertension or diabetes)? (Chapter 3)

3. Do mice with targeted deletion of interleukin 10 gene (IL-10^{tm1Cgn}) have decreased oxygen consumption compared to normal wildtype mice? (Chapter 4)

Our approach to answering these questions was affected by the available data. In fact, a common feature in addressing each of these questions was that information was required from multiple data sources.

Second, we translated each scientific question into an inferential problem involving an appropriate statistical parameter. Third, we investigated what can be learned about the parameter of interest from information available from the observed data. Often this information was not sufficient to learn about the true value of the parameter and additional untestable "identification" assumptions were required. It is important that these assumptions be developed in close collaboration with subject matter experts in order to judge their plausibility. While these assumptions are sufficient to learn about the true value of the parameter in an infinite data setting, we also needed to make additional testable assumptions to ensure that inferences in the finite data setting are reasonably precise.

Fourth, we developed strategies for estimation of the parameters and characterizing their uncertainties. This step involved development of novel inferential methods that combine information from multiple data sources in a statistically principled way. The parameter estimates and associated uncertainties are then used to statistically answer the scientific questions. In particular, the data (plus assumptions) may or may not provide an affirmative answer to the questions. Either way, the result may lead to new questions or theories, which will ideally lead to new discoveries.

In summary, statistical inference procedures using multiple data sources have enormous potential within the scientific discovery process. We believe that the ideas developed in this thesis have broad applicability to other biomedical and public health investigations.

Bibliography

Behseta, S. and Kass, R. E. (2005). Testing equality of two functions using bars. *Statistics in Medicine* **24**(22), 3523–3534.

Behseta, S., Kass, R. E., Moorman, D. E. and Olson, C. R. (2007). Testing equality of several functions: Analysis of single-unit firing-rate curves across multiple experimental conditions. *Statistics in Medicine* **26**(21), 3958–3975.

Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components. *Annals of Statistics* **37**(1), 1–34.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21**(2), 263–275.

Boruvka, A. and Cook, R. J. (2015). A Cox-Aalen model for interval-censored data. *Scandinavian Journal of Statistics* **42**(2), 414–426.

Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D., Thompson, S. G.

and Consortium, EPIC-InterAct. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology* **50**, 543–552.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**(3), 570–579.

Chambers, J. M. and Hastie, T. J. (1991). *Statistical Models in S*. Boca Raton, FL: CRC Press, Inc.

Chen, H. Y., Kittles, R. and Zhang, W. (2013). Bias correction to secondary trait analysis with case-control design. *Statistics in Medicine* **32**(9), 1494–1508.

Chiou, S.H., Kang, S. and Yan, J. (2014). Fitting accelerated failure time models in routine survival analysis with R package aftgee. *Journal of Statistical Software* **61**(11), 1–23.

Corvalan, A., Melo, E., Sherman, R. and Shum, M. (2015). Bounding causal effects in ecological inference problems.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), pp. 187–220.

Crainiceanu, C. M., Staicu, A., Ray, S. and Punjabi, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine* **31**(26), 3223–3240.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L. and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *The Journal of the American Medical Association* **295**(10), 1127–1134.

Eilers, P. H. C. and Marks, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**(2), 89–121.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**(4), 845–854.

Geneletti, S. and Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. In: P. M. Illari, F. Russo and Williamson, J. (editors), *Causality in the Sciences*. New York, NY: Oxford University Press, pp. 728–749.

Ghosh, A., Wright, F. A. and Zou, F. (2013). Unified analysis of secondary traits in case-control association studies. *Journal of the American Statistical Association* **108**(502), 566–576.

Goetghebeur, E. and Ryan, L. M. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* **56**(4), 1139–1144.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A. and Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* **54**(4), 1498–1507.

Gómez, G., Calle, M. Luz and Oller, R. (2004). Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers* **45**(2), 139–173.

Hall, P. and Keilegom, I. Van. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* **17**(2), 1511–1531.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Boca Raton, FL: CRC Press, Inc.

He, J., Li, H., Edmondson, A.C., Rader, D. J. and Li, M. (2011). A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics*.

Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M. and Robins, J.M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* **14**(7), 477–491.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine* **25**(8), 1323–1339.

Klugman, S. A. and Parsa, A. R. (1994). Minimum distance estimation of loss distributions. *Proceedings of the Casualty Actuarial Society* **80**, 250.

Kraft, P. (2007). Analyses of genome-wide association scans for additional outcomes. *Epidemiology* **18**(6), 838.

Langley, P. (1987). *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, MA: MIT Press.

Lee, A. J., Mcmurchy, L. and Scott, A. J. (1997). Re-using data from case-control studies. *Statistics in Medicine* **16**(12), 1377–1389.

Li, H., Gail, M. H., Berndt, S. and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology* **34**(5), 427–433.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology* **33**(3), 256–265.

Lin, X., Cai, B., Wang, L. and Zhang, Z. (2015). A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis* **21**(3), 470–490.

Mokry, L. E., Ross, S., Ahmad, O.S., Forgetta, V., Smith, G. D., Leong, A., Greenwood, C. M. T., Thanassoulis, G. and Richards, J. B. (2015). Vitamin D and risk of Multiple Sclerosis: a Mendelian randomization study. *PLOS Medicine* **12**(8), e1001866.

Monsees, G.M., Tamimi, R. M. and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology* **33**(8), 717–728.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P. and Gutstein, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Annals of Applied Statistics* **5**(2A), 894–923.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A. and Coombes, K. R. (2006). Bayesian analysis of mass spectrometry proteomic data using wavelet based functional mixed models. *Biometrics* **64**(2), 479–489.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(2), 179–199.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association* **98**(463), 573–583.

Nagelkerke, N. J. D., Moses, S., Plummer, F. A., Brunham, R.C. and Fish, D. (1995). Logistic regression in case-control studies: the effect of using independent as dependent variables. *Statistics in Medicine* **14**(8), 769–775.

NCHS, National Center For Health Statistics. (2011). Data file documentation, National Health Interview Survey 2010.

Oller, R., , Gómez, G. and Calle, M. Luz. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics* **32**(3), 315–326.

Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *Journal of Computational and Graphical Statistics* **8**(1), 109–120.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**(1), 199–203.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**(3), 403–411.

Reilly, M., Torrång, A. and Klint, Å. (2005). Re-use of case-control data for analysis of new outcome variables. *Statistics in Medicine* **24**(24), 4009–4019.

Richardson, D. B., Rzehak, P., Klenk, J. and Weiland, S. K. (2007). Analyses of case-control data for additional outcomes. *Epidemiology* **18**(4), 441–445.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83**(2), 355–370.

Satten, G. A., Datta, S. and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93**(441), 318–327.

Sinha, D., Chen, M. and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55**(2), 585–590.

Speakman, J. R. (2013). Measuring energy metabolism in the mouse–theoretical, practical, and analytical considerations. *Frontiers in Physiology* **4**(34), 1–23.

Tchetgen, E. J. Tchetgen. (2014). A general regression framework for a secondary outcome in case–control studies. *Biostatistics* **15**(1), 117–128.

Thakur, N., Oh, S. S., Nguyen, E. A., Martin, M., Roth, L. A., Galanter, J., Gignoux, C. R., Eng, C., Davis, A., Meade, K., LeNoir, M. A., Avila, P. C., Farber, H. J., Serebrisky, D., Brigino-Buenaventura, E., Rodriguez-Cintron, W., Kumar, R., L. K. Williams, K. Bibbins-Domingo, Thyne, S., Sen, S., Rodriguez-Santana, J. R., Borrell, L. N. *and others.* (2013). Socioeconomic status and childhood asthma in urban minority youths. the GALA II and SAGE II studies. *American Journal of Respiratory and Critical Care Medicine* **188**(10), 1202–1209.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(3), 290–295.

Valappil, T. I., Singh, K. P. and Bartolucci, A. A. (2015). An overview of estimating causal effects from interval-censored data: G-estimation approach. *In preparation*.

Vandebosch, A. and Goetghebeur, E. (2005). Structural proportional hazards models for the effects of observed exposures based on interval censored survival data in randomized clinical trials. *In preparation*.

Wang, J. and Shete, S. (2011). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology* **35**(3), 190–200.

Wang, L., McMahan, C. S., Hudgens, M. G. and Qureshi, Z. P. (2015). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 10.1111/biom.12389.

Wang, X., Chen, M. and Yan, J. (2013). Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Analysis* **19**(3), 297–316.

Wei, J., Carroll, R. J., Müller, U. U., Keilegom, I. V. and Chatterjee, N. (2013). Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1), 185–206.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* **11**(14-15), 1871–1879.

Yavuz, A. and Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine* **30**(1), 75–90.

Zhang, Y., Hua, L. and Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**(2), 338–354.

Vita

Personal Data

Place of Birth:	Kolkata, India
Date of Birth:	12/31/1986
Address:	615 North Wolfe Street, E3001 BSPH Baltimore, MD 21205-2179, USA.
Phone:	+1 (443) 287-4788 / +1 (443) 813-1983
email:	parichoy@jhu.edu / ppcparichoy@gmail.com

Research Experience

Aug 2010 – Biostatistics PhD Candidate

Advisor: Dr. Daniel O. Scharfstein

Department of Biostatistics, The Johns Hopkins University

My research is focussed on learning about scientific and causal questions from multiple data sources in scenarios where a single data source is not sufficient to answer the question of interest. I am broadly interested in causal inference and have been working on developing methods and tools aimed to solve problems arising in applications with real data from genetic epidemiology, environmental epidemiology, clinical trials and electronic medical records.

Positions Held

Sept 2011 – Graduate Research Assistant
Department of Biostatistics, The Johns Hopkins University
Mentor: Dr. Daniel O. Scharfstein (and others)

Jan 2013 – Graduate Research Assistant and Statistical Consultant
Center of Aging and Health, The Johns Hopkins University
Mentor: Drs. Qian-li Xue, Karen Bandeen Roche (and others)

Research Interests

Causal Inference, Case-control Studies, Survival Analysis, Functional Data Analysis, Clinical Trials, Semiparametrics and Missing Data, Statistical Genetics, Applied Statistics

VITA

Education

Aug 2010 –	PhD, Biostatistics (Expected: Jan 2016) **The Johns Hopkins University** Baltimore, MD USA **Thesis:** Statistical Inference with Multiple Data Sources
Jul 2008 – May 2010	M.Stat., Statistics (First Division with Distinction) **Indian Statistical Institute** Kolkata, WB, India **Thesis:** A Study of Association between a Quantitative Trait and Two Locus Marker Haplotypes
Jul 2005 – May 2008	B.Stat.(Hons.), Statistics (First Division with Distinction) **Indian Statistical Institute** Kolkata, WB, India

Software and Computer Skills

R, MATLAB, SAS, Latex, MS Office

VITA

Publications

1. **Pal Choudhury, P.**, Scharfstein, D. O., Diaz, I., Mcmahan, C., Luo, X., Massie A.B., Segev, D.L. (2016). Causal effect among the exposed: multiple data sources and censored outcomes. *In preparation.*

2. **Pal Choudhury, P.**, Scharfstein, D. O., Galanter, J. M., Gignoux, C. R., Roth, L. A., Oh, S. S., Borrell, L. N., Burchard, E. G., Sen, S. (2016). Enhancing genetic case-control studies using sample surveys. *In preparation.*

3. **Pal Choudhury, P.**, Crainiceanu, C., Westbrook, R., Xue QL. (2016). Testing equality of curves after covariate adjustment. *In preparation.*

4. Westbrook, R., Langdon, J.M., Roy, C.N., Yang, H., **Pal Choudhury, P.**, Xue, QL., de Cabo, R., Walston, J. (2016). The metabolic characterization of a frail mouse model: matching statistical methods to analytic objectives. *In preparation.*

5. Psoter, K. J., Diaz, I., **Pal Choudhury, P.**, Rosenfeld, M., Carone, M., Scharfstein, D. O. (2016). Estimating the causal effect of a point exposure: MRSA infection and subsequent initial *Pseudomonas aeruginosa* acquisition in young children with cystic fibrosis. *In preparation.*

6. Buta, B., **Pal Choudhury, P.**, Xue, QL., Chaves, P., Bandeen-Roche, K., Walston, J., Semba, R., Shardell, M., Michos, E., Ferrucci, L., Gross, A., McAdams,

M., Kalyani, R. (2016). Vitamin D, cardiometabolic diseases, and the incidence of frailty in older women. *In preparation.*

7. **Pal Choudhury, P.**, Bagchi, P. , Sengupta S., Ghosh A. (2010). On effect of compromised nodes on security of wireless sensor network. Ad Hoc Sensor Wireless Networks 9, 255-273.

Conferences and Seminars

1. Jan 2016: **Pal Choudhury, P.** *Statistical inference with multiple data sources.*

   Thesis Defense Seminar: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

2. Oct 2015: **Pal Choudhury, P.**, Scharfstein, D. O., McMahan, C., Massie, A., Segev, D. *Causal effect among the exposed: multiple data sources and censored outcomes.*

   Transplant Epidemiology Research Group Meeting: Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

3. Mar 2015: **Pal Choudhury, P.**. *The sign of the logistic regression coefficient.*

   Biostatistics Journal Club: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

4. Mar 2015: **Pal Choudhury, P.**, Scharfstein, D. O. *On causal inference about treatment effect in studies with randomized and observational components.*

   Causal Inference Seminar: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

5. Jan 2015: **Pal Choudhury, P.**, Xue, Q. L. , Westbrook, R. *The metabolic characterization of a frail mouse model: matching statistical methods to analytic objectives.*

   Biostatistics Seminar: Center of Aging and Health, The Johns Hopkins University, Baltimore, MD, USA.

6. Aug 2014: **Pal Choudhury, P.**, Scharfstein, D. O., Sen, S. *Enhancing genetic case-control studies using sample surveys.*

   Contributed Talk: JSM 2014. Boston, MA, USA

7. Apr 2014: **Pal Choudhury, P.**. *Mendelian randomization: a review from a causal inference perspective.*

   Biostatistics Journal Club: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

8. Apr 2014: **Pal Choudhury, P.**, Scharfstein, D. O., Sen, S. *Enhancing genetic case-control studies using sample surveys.*

   Student Paper Competition: Probability and Statistics Day 2014, University of Maryland, Baltimore County, Arbutus, MD, USA.

9. Mar 2014: **Pal Choudhury, P.**, Scharfstein, D. O., Sen, S. *Enhancing genetic case-control studies using sample surveys.*

   Contributed Talk: ENAR 2014. Baltimore, MD, USA

10. Jan 2014: **Pal Choudhury, P.**, Scharfstein, D. O. *Mendelian randomization: a review from a causal inference perspective.*

    Invited Talk: Applied Statistics Unit Seminar, Indian Statistical Institute, Kolkata, India

11. Dec 2013: **Pal Choudhury, P.**, Scharfstein, D. O., Schwartz, B. *Analyzing the Causal Effect of Cumulative Lead Dose on Cognitive Function.*

    Causal Inference Seminar: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

12. Oct 2013: **Pal Choudhury, P.**, Scharfstein, D. O., Sen, S. *Enhancing genetic case-control studies using sample surveys.*

    Causal Inference Seminar: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

13. Jan 2013: **Pal Choudhury, P.**, Scharfstein, D. O. *On analysis of studies with missing venography data.*

    Invited Talk: Applied Statistics Unit Seminar, Indian Statistical Institute, Kolkata, India.

14. Dec 2012: **Pal Choudhury, P.**, Scharfstein, D. O., Sen, S. *Enhancing genetic*

case-control studies using sample surveys.

Contributed Talk: Eighth International Triennial Calcutta Symposium on Probability and Statistics, University of Calcutta, Kolkata, India.

15. Jun 2009: **Pal Choudhury, P.** and others. *Analysis of women drop-out rate in India.*

    Project Presentation: Department of Higher Education, Ministry of Human Resource and Development, Govt. of India, New Delhi, India.

16. Aug 2008: **Pal Choudhury, P.**, Bagchi, P. , Sengupta, S., Ghosh, A. *On effect of compromised nodes on security of wireless sensor network.*

    Invited Talk: National Workshop on Cryptology 2008, University of Hyderabad, Hyderabad, India.

17. Jul 2008: **Pal Choudhury, P.**, Chakravarti, A. *Sampling rare alleles by sequencing individuals and populations.*

    Project Presentation: Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

18. Dec 2007: **Pal Choudhury, P.**, Bambardekar, K. *Biology of malaria parasites: plasmodium infected red blood cells under the action of optical tweezers.*

    Project Presentation: Tata Institute of Fundamental Research, Mumbai, India

VITA

Memberships

- American Statistical Association (ASA)

- International Indian Statistical Association (IISA)

Honors and Awards

- **2014:** Second Best Paper Award in the Graduate Students Oral Presentations at 8th Annual Probability and Statistics Day at University of Maryland Baltimore County.

- **2012:** Joseph Zeger Travel Award for presenting the paper "Enhancing Genetic Case-Control Studies Using Sample Surveys" at the Eighth International Triennial Calcutta Symposium on Probability and Statistics, Kolkata, India.

- **2008**: Summer Travel Awards from Sir Dorabji Tata Trust, Mumbai, India for traveling to the Johns Hopkins University, Baltimore, MD, USA and pursue research internship in Statistical Genetics.

- **2005**: M. P. Birla Foundation Award for bagging 4^{th} place in my school (among 500 students) and 13^{th} place in my state (among 394,636 students) in the school leaving examinations (Absolute aggregate score percentage: 95.4).

VITA

Parichoy Pal Choudhury is starting as a Postdoctoral Research Fellow in Biostatistics with Dr. Nilanjan Chatterjee (Bloomberg Distinguished Professor of Biostatistics and Oncology) at the Johns Hopkins University from January 2016. He will be working at the interface of causal inference, statistical genetics and disease risk modeling.