CAUSAL INFERENCE BEYOND ESTIMATING AVERAGE TREATMENT EFFECTS

Kwonsang Lee

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Graduate Group Chairperson

Charles L. Epstein, Thomas A. Scott Professor of Mathematics

Dissertation Committee

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Paul R. Rosenbaum, Robert G. Putzel Professor of Statistics

Bhaswar B. Bhattacharya, Assistant Professor of Statistics

CAUSAL INFERENCE BEYOND ESTIMATING AVERAGE TREATMENT EFFECTS

Kwon Sang Lee

ACKNOWLEDGEMENT

First, I would like to thank my advisor, Dr. Small. His sharp intuition, brilliant ideas and endless enthusiasm for pursuing research make him the ideal researcher I aspire to be. His extensive knowledge inside and even outside of statistical literature has constantly been a great influence on my research during my doctoral program. I was very fortunate to have had the opportunity to start a project with him in my second year. Under his kind guidance, I have learned so much about causal inference and had many opportunities to collaborate and interact with other researchers. All the experiences that he provided created a solid foundation for my current thesis on causal inference.

Second, I want to thank my dissertation committee members, Dr. Rosenbaum and Dr. Bhattacharya. The Wednesday meetings that I usually had with Dr. Rosenbaum were intellectually inspiring, and helped me cultivate a deeper understanding of effect modification and sensitivity analysis in observational studies. I also want to thank Dr. Bhattacharya for his ingenious ideas that offered a significant breakthrough when proving the asymptotic properties of the nonparametric binomial likelihood estimator. I always enjoyed every discussion with him and truly appreciated his keen and remarkable insight.

Finally, I want to thank my family for supporting me during my Ph.D. years. I cannot thank my parents enough for their unwavering support throughout the years. Even from a distance, their phone calls and texts were great encouragement to me. Above all, I would not have been able to even start this Ph.D. program and graduate it without the support of my wife, Stella. She was always on my side, reassured me when I was frustrated, and helped me get through so many of the challenges I faced. Every moment of my Ph.D. journey was joyful, even through tough times, because I had her by my side.

ABSTRACT

CAUSAL INFERENCE BEYOND ESTIMATING AVERAGE TREATMENT EFFECTS

Kwonsang Lee

Dylan S. Small

Many scientific questions are to understand and reveal the causal mechanisms from observational study data or experimental data. Over the past several decades, there has been a large number of developments to render causal inferences from observed data. Most developments are designed to estimate the mean difference between treated and control groups that is often called the average treatment effect (ATE), and rely on identifying assumptions to allow causal interpretation. However, more specific treatment effects beyond the ATE can be estimated under the same assumptions. For example, instead of estimating the mean of potential outcomes in a group, we may want to estimate the distribution of the potential outcomes. Understanding the distribution implies understanding the mean, but not vice versa. Therefore, more sophisticated causal inference can be made from the data. The dissertation focuses on causal inference in observational studies, and discusses three main achievements. First, in instrumental variable (IV) models, we propose a novel nonparametric likelihood method for estimating the distributional treatment effect that compares two potential outcome distributions for treated and control groups. Furthermore, we provide a nonparametric likelihood ratio test for the hypothesis that the two potential outcome distributions are identical. Second, we develop two methods for discovering effect modification in a matched observational study data: (1) the CART method and (2) the Submax method. Both methods are applied to real data examples for finding effect modifiers that alter the magnitude of the treatment effect. Lastly, we provide a causal definition of the malaria attributable fever fraction (MAFF) that has not been studied in the causal inference field, and propose a novel maximum likelihood method to account for fever killing effect and measurement errors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

One common fallacy is the belief that two correlated variables indicate a cause-and-effect relationship, thus the oft-cited warning: correlation does not imply causation. This warning is often illustrated with examples of false causation such as living near overhead power lines causes cancer in children such as leukemia Wertheimer and Leeper (1979). An important confounding factor that misled the researchers to infer a causal relationship was the factor of income. Living under power lines is often a low-income housing location and there is a strong, well-known epidemiological relationship between poverty and cancer Aber et al. (1997); Bona et al. (2016). This research incited public panic, and countless public health efforts and costs were spent to assuage people of their fear of power lines. Without proper methodological arguments, nothing can be said about causality from correlated data. Causal inference is the field of study of how and to what extent we can infer causality from data.

In causal inference, the effect of a treatment on an outcome can be identified for either randomized experiments or observational study designs. The gold standard for investigating causal treatment effects is to conduct a randomized experiment. Randomized experiments randomly assign individuals to treatment or control, ensuring that the treated and control groups are comparable such that differences in outcomes between the groups can be attributed to the treatment. Randomized experiments are useful, but they are not always ethical or feasible. Alternatively, we may attempt to draw causal inferences from observational studies. Unlike randomized experiments, in observational studies, researchers have no control of treatment assignment. This distinction brings many additional challenges. For example, differences in outcomes between a treated group and a control group may reflect differences of covariates between the two groups rather than the effect of a treatment. Two popular methods for drawing causal inferences from observational data are instrumental variable (IV) Imbens and Angrist (1994a); Angrist et al. (1996) and matching methods Rosenbaum (2002b); both can infer the effect of a treatment by adjusting for confounding covariates.

Most literature has focused on identifying the average treatment effect (ATE), which is the expected benefit from treatment on average in the population. However, from the same study design, further sophisticated inferences can be made: (1) estimating the effect of a treatment on the distribution of outcomes, *the distributional treatment effect* and (2) discovering which subgroup of subjects has the largest treatment effect, *effect modification*. First, the distributional treatment effect goes beyond the ATE by providing a comprehensive understanding of the impact of the treatment on the whole population. For instance, if researchers want to study how a new policy designed to decrease income inequality affects household income, they cannot use average income as a measure. Rather, the whole income distribution must be examined to see if the new policy has an effect on the growth of the middle class. Second, discovering effect modification provides inferences that are less sensitive to unmeasured confounding, which is important when conducting sensitivity analysis in observational studies. Besides this theoretical advantage, discovering effect modification has practical implications. For example, if doctors recognize which patients can benefit more from taking a treatment, then they can use this information to develop personalized treatments Coalition PM (2014); Hamburg and Collins (2010). Inferences about distribu-

tional treatment effects and effect modification enable researchers to evaluate the effect of a treatment in more accurate and diverse ways .

Furthermore, causal inference can be more generally used to define other measures with causal interpretation. Most of the time, the effect of treatment on outcome is of interest, and the target estimand is the difference of the average potential outcome between treated and control. However, other estimands may be of interest in many fields such as epidemiology, medicine and public health. For example, the malaria attributable fever fraction (MAFF), i.e., the proportion of fevers that are attributable to (caused by) malaria parasites, is an important public health measure for assessing the effect of malaria control programs and other purposes. The MAFF provides information about the public health burden from the malaria, and how much resources should be devoted to combatting malaria compared to other diseases.

The remainder of this thesis is organized as follows. We begin in Chapter 2 with estimating the distributional treatment effect in IV models. Next in Chapter 3 and 4, we introduce two approaches for discovering effect modification in observational studies with matching. One proposes a novel approach based on classification and regression cart, which is one of well-known machine learning techniques, and the other proposes a novel statistical approach based on multiple testing correction by providing tractable statistical properties. In Chapter 5, we propose a novel maximum likelihood estimation method based on g-modeling to estimate the MAFF in the potential outcome framework by accounting for fever killing effects and measurement errors. Lastly, we provide discussions in Chapter 6.

CHAPTER 2 : Nonparametric Inference for Distributional Treatment Effects in Instrumental Variable Models

2.1. Introduction

Randomized experiments are the gold standard for assessing the effect of a treatment but often it is not practical or ethical to randomly assign a treatment itself. However, in some settings, an encouragement to take the treatment can be randomized Holland (1988). In other settings, no randomization is possible but there may be a "natural experiment" such that some people are encouraged to receive the treatment compared to others in a way that is effectively random Angrist and Krueger (2001). For both of these settings, the instrumental variable (IV) method can be used to estimate the causal effect of a treatment Holland (1988); Angrist et al. (1996). The IV method is a method that controls for unmeasured confounders to make causal inferences about the effect of a treatment. An IV is informally a variable that affects the treatment but is independent of unmeasured confounders and only affects the outcome through affecting the treatment (see Section 2.2 for a more precise definition). Under a monotonicity assumption that the encouraging level of the IV never causes someone not to take the treatment, the IV method identifies the treatment effect for compliers, those subjects who would take the treatment if they received the encouraging level of the IV but would not take the treatment if they did not receive the encouraging level Angrist et al. (1996). For several discussions of the IV method, see Abadie (2003), Angrist et al. (1996), Baiocchi et al. (2014), Brookhart and Schneeweiss (2007), Cheng et al. (2009), Hernan and Robins (2006), Ogburn et al. (2015) and Tan (2006).

Much of the literature on the treatment effect in instrumental variable models has focused on estimating the average treatment effect for compliers. However, understanding the effect of the treatment on the whole distribution of outcomes for the compliers, *the distributional treatment effect* for compliers, is important for optimal individual decision-making and for social welfare comparisons. Optimal individual decision-making requires computing the expected utility of the treatments which requires knowing the whole distribution of the outcomes under the treatments being compared rather than just the average outcomes when the utility function is nonlinear Karni (2009). Social welfare comparisons require integration of utility functions under the distribution of the outcome (say income), which again requires knowing the effect of the treatment on the whole distribution of outcomes Abadie (2002); Atkinson (1970).

Abadie (2002) developed a nonparametric method for estimating the effect of treatment on the cumulative distribution function (CDF) of the outcomes based on expanding the conventional IV approach described in Imbens and Angrist (1994b). The estimator is based on a method of moments (MOM) approach. Though the MOM method identifies the CDFs of compliers, the MOM estimate can violate the conditions of a CDF such as nondecreasingness and nonnegativeness. In this paper, we develop a nonparametric likelihood based approach that enforces that the CDF estimates should be nondecreasing and nonnegative. Nonparametric likelihood methods have been shown to have appealing properties in many settings such as providing nonparametric inferences that inherit some of the attractive properties of parametric likelihood (e.g., automatic determination of the shape of confidence regions)

and straightforward interpretation of side information expressed through constraints Owen (2001). However, the usual nonparametric likelihood approach does not work for the IV model because there are infinitely many solutions that maximize the likelihood. In this case, the usual nonparametric likelihood method fails to produce a meaningful estimator Geman and Hwang (1982).

We propose a novel adaptation of nonparametric likelihood that overcomes the problem of usual nonparametric likelihood for our setting. Our approach builds on the fact that MOM method identifies the CDFs for compliers at any given point using information on whether outcomes are less than or equal to vs. above that point, in other words based on binomially distributed random variables. We consider the composite likelihood that multiplies together the pieces of the likelihood contributed by these binomial random variables. This is a composite or "pseudo" likelihood rather than a true likelihood because the binomial random variables are actually dependent but are treated as independent in the composite likelihood. Composite likelihood has been found useful in a range of areas including problems in geostatistics, spatial extremes, space-time models, clustered data, longitudinal data, time series and statistical genetics Lindsay (1988); Heagerty and Lele (1998); Varin et al. (2011); Larribe and Fearnhead (2011). We call the composite likelihood method that we use in our setting the maximum binomial likelihood (MBL) method because it maximizes the average of the likelihood of the binomial random variables at each point across all observation points. We develop a computationally fast method for finding the MBL estimate by combining the expectation maximization (EM) and pool adjacent violators algorithm (PAVA). Unlike the usual nonparametric maximum likelihood, maximizing the binomial likelihood produces a unique estimate. We show that the MBL estimator is consistent and demonstrate in simulation studies that the MBL estimator performs better than other estimators, particularly when the IV is weak (weakly associated with the treatment). The advantage of the maximum binomial likelihood method over the nonparametric MOM method is that it ties together the information used by the nonparametric MOM method to identify the CDFs for compliers at each observation point rather than treating them separately and enforces the constraints that CDFs are nonnegative and nondecreasing.

The MBL method can be used to construct a hypothesis test of no distributional treatment effect by using a binomial likelihood ratio test statistic. A test of no treatment effect was previously discussed based on the Kolmogorov-Smirnov test statistic in Abadie (2002). This approach does not bring in the structure of the IV model. Our approach uses the structure of the IV model and can be much more powerful.

We apply the MBL method to a study of the effect of participation in the student breakfast program (SBP) on childhood obesity. Participation in SBP is not randomized and counfounders may exist in the relationship between participation in the SBP and obesity. To control for possible unmeasured confounders, we propose to use the variable of the distance from children's homes to their schools as an IV. If a child lives near schools, then she is more likely to participate the program. We assume that this IV is randomized. The validity of this assumption is discussed in Section 2.5. Using this binary distance variable, we make inferences about the distributional effect of SBP participation on childhood obesity from the Early Childhood Longitudinal Program - Kindergarten Class (ECLS-K) 2010-2011 data.

Table 1: Compliance classes by the potential outcomes $D_i(0)$ and $D_i(1)$

	$D_i(0) = 0$	$D_i(0) = 1$
$D_i(1) = 0$	Never-takers	Defiers
$D_i(1) = 1$	Compliers	Always-takers

The rest of this article is organized as follows. In Section 2.2, basic notation and assumptions in instrumental variable models using the potential outcome framework are introduced. Also, the existing nonparametric method, the MOM method, is reviewed. The motivation for a new approach of constructing nonparametric likelihood is provided along with the difficulty of applying the usual nonparametric likelihood methods to IV models. In Section 2.3, we develop the MBL method for estimating the CDFs based on the binomial likelihood approach. Simulation studies are conducted to assess the performance of the proposed estimation method. Section 2.4 describes a new test statistic based on the binomial likelihood with simulation studies. In Section 2.5, the MBL method and the MOM method are applied to the veterans data, and compared to one another. Section 2.6 includes summary.

2.2. Framework and Review

In this section the framework of instrumental variable (IV) model is introduced and the existing methods are briefly reviewed. Notation and identification assumptions are discussed in Section 2.2.1. The existing method of moments approach (Abadie, 2002) for estimating distributional treatment effects in IV models is reviewed, and the shortcomings of this method are addressed in Section 2.2.2. This motivates our new approach of constructing the nonparametric binomial likelihood, introduced in Section 2.3.

2.2.1. Notation and Assumption

Let Z_a be a binary instrumental variable and D_a be an indicator variable for whether subject $a \in [n] := \{1, 2, \ldots, n\}$ receives the treatment or not. Using the potential outcome framework, define $D_a(0)$ as the value that D_a would be if Z_a were to be set to 0, and $D_a(1)$ as the value that D_a would be if Z_a were to be set to 1. Similarly, $Y_a(z, d)$ for $(z, d) \in \{(0,0), (0,1), (1,0), (1,1)\}$, is the value that the outcome Y_a would be if Z_a were to be set to z and D_a were to be set to d. For each subject $a \in [n]$, the analyst can only observe one of the two potential values $D_a(0)$ and $D_a(1)$, and one of the four potential values $Y(0,0), Y(0,1), Y(1,0), Y(1,1)$. The observed treatment D_a is

$$D_a = Z_a D_a(1) + (1 - Z_a) D_a(0).$$

Similarly, the observed outcome Y_a can be expressed as $Y_a = Z_a D_a \cdot Y_a(1,1) + Z_a(1 - D_a) \cdot Y_a(1,0) + (1 - Z_a) D_a \cdot Y_a(0,1) + (1 - Z_a)(1 - D_a) \cdot Y_a(0,0)$. A subject's *compliance class* is determined by the combination of the potential treatment values $D_a(0)$ and $D_a(1)$, which is denoted by S_a: $S_a =$ *always-taker* (at) if $D_a(0) = 1, D_a(1) = 1$; $S_a =$ *never-taker* (nt) if $D_a(0) = 0, D_a(1) = 0$; $S_a =$ *complier* (co) if $D_a(0) = 0, D_a(1) = 1$; and $S_a =$ *defier* (de) if $D_a(0) = 1, D_a(1) = 0$. This is summarized in Table 1.

For the rest of this article, the following standard identifying conditions are assumed. The

implications of these conditions are briefly explained in the paragraph below (refer to Angrist et al. (1996) for more details on these assumptions).

Assumption 1. Hereafter, the following identification conditions will be imposed on the IV model:

(1) *Stable Unit Treatment Value Assumption* (SUTVA) Rubin (1986): The outcome (treatment) for individual $i \in [n]$ is not affected by the values of the treatment or instrument (instrument) for other individuals and that the outcome (treatment) does not depend on the way the treatment or instrument (instrument) is administered.

(2) *The instrumental variable Z_a is independent of the potential outcomes $Y_a(z,d)$ and potential treatment $D_a(z)$.*

$$Z_a \perp\!\!\!\perp (Y_a(0,0), Y_a(0,1), Y_a(1,0), Y_a(1,1), D_a(0), D_a(1))$$

(3) *Nonzero average causal effect of Z on D*: $\mathbb{P}(D_a(1)=1) > \mathbb{P}(D_a(0)=1)$.

(4) *Monotonicity*: $D_a(1) \geq D_a(0)$.

(5) *Exclusion restriction*: $Y_a(0,d) = Y_a(1,d)$, for $d = 0$ or 1.

Assumption 1 enables the causal effect of the treatment for the subpopulation of the compliers to be identified. The SUTVA allows us to use the notation $Y_a(z,d)$ (or $D_a(z)$), which means that the outcome (treatment) for individual i is not affected by the values of the treatment and instrument (instrument) for other individuals. Condition (2) will be satisfied if Z_a is randomized. Condition (3) requires Z to have some effect on the average probability of treatment. Condition (4), the monotonicity assumption, means that the possibility of $D_a(0) = 1$, $D_a(1) = 0$ is excluded, that is, there are no defiers (see Table 1). Condition (5) assures that any effect of Z on Y must be through an effect of Z on D. Under this assumption, the potential outcome can be written as $Y_a(d)$, instead of $Y_a(z,d)$.

Distribution Functions of the Compliance Classes

Define the distribution functions of compliers without treatment, never-takers, compliers with treatment, and always-takers respectively:

$$\begin{aligned}
F_{co}^{(0)}(t) &= \mathbb{E}[\mathbf{1}\{Y_1(0) \leq t\} | D_1(1) = 1, D_1(0) = 0], \\
F_{nt}(t) &= \mathbb{E}[\mathbf{1}\{Y_1(0) \leq t\} | D_1(1) = 0, D_1(0) = 0], \\
F_{co}^{(1)}(t) &= \mathbb{E}[\mathbf{1}\{Y_1(1) \leq t\} | D_1(1) = 1, D_1(0) = 0], \\
F_{at}(t) &= \mathbb{E}[\mathbf{1}\{Y_1(1) \leq t\} | D_1(1) = 1, D_1(0) = 1].
\end{aligned} \tag{2.2.1}$$

Under Assumption 1, the distributions are identified such as

$$\begin{aligned} F_{co}^{(0)}(t) =& \mathbb{P}(Y_1 \le t | Z_1 = 0, S_1 = co), \\ F_{nt}(t) =& \mathbb{P}(Y_1 \le t | Z_1 = 0, S_1 = nt), \\ F_{co}^{(1)}(t) =& \mathbb{P}(Y_1 \le t | Z_1 = 1, S_1 = co), \\ F_{at}(t) =& \mathbb{P}(Y_1 \le t | Z_1 = 1, S_1 = at). \end{aligned} \tag{2.2.2}$$

Moreover, for $u, v \in \{0, 1\}$, define

$$F_{uv}(t) = \mathbb{P}(Y_1 \le t | Z_1 = u, D_1 = v). \tag{2.2.3}$$

Note that

$$\begin{aligned} F_{00}(t) &= \mathbb{P}(Y_1 \le t | Z_1 = 0, D_1 = 0) \\ &= \frac{\mathbb{P}(Y_1 \le t, Z_1 = 0, D_1 = 0, S_1 = co) + \mathbb{P}(Y_1 \le t, Z_1 = 0, D_1 = 0, S_1 = nt)}{\mathbb{P}(Z_1 = 0, D_1 = 0)} \\ &= \lambda_0 \mathbb{P}(Y_1 \le t | Z_1 = 0, S_1 = co) + (1 - \lambda_0) \mathbb{P}(Y_1 \le t | Z_1 = 0, S_1 = nt) \\ &= \lambda_0 F_{co}^{(0)}(t) + (1 - \lambda_0) F_{nt}(t). \end{aligned} \tag{2.2.4}$$

where $\lambda_0 = \mathbb{P}(S_1 = co | Z_1 = 0, D_1 = 0)$. Similarly, it follows that $F_{01}(t) = F_{at}(t)$, $F_{10}(t) = F_{nt}(t)$, and

$$F_{11}(t) = \lambda_1 F_{co}^{(1)}(t) + (1 - \lambda_1) F_{at}(t). \tag{2.2.5}$$

where $\lambda_1 = \mathbb{P}(S_1 = co | Z_1 = 1, D_1 = 1)$.

Next, consider the unknown population proportions of compliance classes $\phi_{co} = \mathbb{P}(S_1 = co), \phi_{at} = \mathbb{P}(S_1 = at), \phi_{nt} = \mathbb{P}(S_1 = nt)$, with $\phi_{co} + \phi_{at} + \phi_{nt} = 1$. Then

$$\lambda_0 = \frac{\phi_{co}}{\phi_{co} + \phi_{nt}}, \quad \lambda_1 = \frac{\phi_{co}}{\phi_{co} + \phi_{at}}. \tag{2.2.6}$$

For $u, v \in \{0, 1\}$, define $n_{uv} := \sum_{a=1}^{n} \mathbf{1}\{Z_a = u, D_a = v\}$. Assume that there exists $\eta_{uv} > 0$ such that $\lim_{n \to \infty} n_{uv}/n \to \eta_{uv} > 0$, for all $u, v \in \{0, 1\}$. Then, from Assumption 1,

$$\eta_{00} = \phi_0(\phi_{co} + \phi_{nt}), \ \eta_{01} = \phi_0 \phi_{at}, \ \eta_{10} = \phi_1 \phi_{nt}, \ \eta_{11} = \phi_1(\phi_{co} + \phi_{at}) \tag{2.2.7}$$

where $\phi_0 = \mathbb{P}(Z = 0)$ and $\phi_1 = 1 - \phi_0 = \mathbb{P}(Z = 1)$. Note that ϕ_0, ϕ_1 and $\eta_{00}, \eta_{01}, \eta_{10}, \eta_{11}$ (hence λ_0 and λ_1) can be estimated directly from the sample proportions as follows:

$$\breve{\phi}_0 = \frac{n_{00} + n_{01}}{n}, \quad \breve{\phi}_n = \frac{n_{10}}{n_{10} + n_{11}}, \quad \breve{\phi}_a = \frac{n_{01}}{n_{00} + n_{01}}, \quad \breve{\phi}_c = 1 - \breve{\phi}_n - \breve{\phi}_a \tag{2.2.8}$$

These estimates will be referred to as the *plug-in estimates*.

Finally, for $u, v \in \{0,1\}$, denote the empirical analogues of (2.2.3) as follows:

$$\mathbb{F}_{uv}(t) = \frac{1}{n_{uv}} \sum_{a=1}^{n} \mathbf{1}\{Z_a = u, D_a = v, Y_a \leq t\}. \tag{2.2.9}$$

Observe that the empirical distribution function $\mathbb{H}$ of the overall data as

$$\mathbb{H}(t) := \frac{1}{n} \sum_{a=1}^{n} \mathbf{1}\{Y_a \leq t\} = \sum_{u,v \in \{0,1\}} \left(\frac{n_{uv}}{n}\right) \mathbb{F}_{uv}. \tag{2.2.10}$$

Note that $\mathbb{H}$ converges uniformly to the limiting distribution $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$, as $n \to \infty$.

Parameter Space

Denote the set of all functions from $\mathbb{R} \to \mathbb{R}$ by $\mathbb{R}^{\mathbb{R}}$, the set of all functions from $\mathbb{R} \to [0,1]$ by $[0,1]^{\mathbb{R}}$ and the set of distribution functions from $\mathbb{R} \to [0,1]$ by $\mathcal{P}([0,1]^{\mathbb{R}})$. Define the *unrestricted* parameter space

$$\boldsymbol{\vartheta} = \left\{(\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at}) : \theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} \in \mathbb{R}^{\mathbb{R}}\right\}. \tag{2.2.11}$$

The *restricted* parameter space is the subset of $\boldsymbol{\vartheta}$ where each $\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at}$ is a distribution function. Formally,

$$\boldsymbol{\vartheta}_+ = \left\{(\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at}) : \theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\right\}. \tag{2.2.12}$$

We note that the MOM estimator lies in $\boldsymbol{\vartheta}$, but it might not be in $\boldsymbol{\vartheta}_+$.

Under the null hypothesis of $F_{co}^{(0)} = F_{co}^{(1)}$ the *restricted null* parameter space is

$$\boldsymbol{\vartheta}_{+,0} = \left\{(\theta_{co}, \theta_{nt}, \theta_{at}) : \theta_{co}, \theta_{nt}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\right\}. \tag{2.2.13}$$

The *unrestricted null* parameter space ϑ_0 can be defined similarly.

2.2.2. Review of The Existing Nonparametric Method (MOM)

Abadie (2002) proved that if $h(\cdot)$ is a measurable function on the real line such that $\mathbb{E}|h(Y_1)| < \infty$ and Assumptions 1 holds, then

$$\begin{aligned}
\frac{\mathbb{E}[h(Y_1)D_1|Z_1=1] - \mathbb{E}[h(Y_1)D_1|Z_1=0]}{\mathbb{E}[D_1|Z_1=1] - \mathbb{E}[D_1|Z_1=0]} &= \mathbb{E}[h(Y_1(1))|D_1(0)=0, D_1(1)=1], \\
\frac{\mathbb{E}[h(Y_1)(1-D_1)|Z_1=1] - \mathbb{E}[h(Y_1)(1-D_1)|Z_1=0]}{\mathbb{E}[(1-D_1)|Z_1=1] - \mathbb{E}[(1-D_1)|Z_1=0]} &= \mathbb{E}[h(Y_1(0))|D_1(0)=0, D_1(1)=1].
\end{aligned}$$

This gives formulas for the CDFs of the potential outcome for compliers under treatment

and control when $h(Y_1)$ is replaced by $\mathbf{1}\{Y_1 \leq t\}$. This gives

$$F_{co}^{(1)}(t) = \frac{\mathbb{E}[\mathbf{1}\{Y_1 \leq t\}D_1|Z_1 = 1] - \mathbb{E}[\mathbf{1}\{Y_1 \leq t\}D_1|Z_1 = 0]}{\mathbb{E}[D_1|Z_1 = 1] - \mathbb{E}[D_1|Z_1 = 0]}, \tag{2.2.14}$$

and

$$F_{co}^{(0)}(t) = \frac{\mathbb{E}[\mathbf{1}\{Y_1 \leq t\}(1 - D_1)|Z_1 = 1] - \mathbb{E}[\mathbf{1}\{Y_1 \leq t\}(1 - D_1)|Z_1 = 0]}{\mathbb{E}[(1 - D_1)|Z_1 = 1] - \mathbb{E}[(1 - D_1)|Z_1 = 0]}. \tag{2.2.15}$$

Abadie (2002) proposed substituting the sample means for the expectation in (2.2.14) and (2.2.15) to estimate the CDFs for the compliers nonparametrically. These are the well-known nonparametric method of moments (MOM) estimates, and will be denote by $\breve{F}_{co}^{(1)}(t)$ and $\breve{F}_{co}^{(0)}(t)$, respectively.

There are three problems with the nonparametric MOM method which this paper seeks to improve:

(1) The nonparametric MOM estimates $\breve{F}_{co}^{(1)}(t)$ and $\breve{F}_{co}^{(0)}(t)$ might violate the non-decreasing condition of distribution functions.

(2) The MOM estimates might produce estimates which are outside of the interval [0,1]. This is called the *violation of non-negativeness*.

(3) Finally, MOM estimates could be highly unstable in the weak instrument setting (meaning that the IV is only weakly associated with the treatment so that there are a small proportion of compliers) because the denominators of both equations (2.2.14) and (2.2.15) depend on the proportion of compliers in the entire population.

The three problems arise at the same time when the IV is weak or the sample size is relatively small. The maximum binomial likelihood (MBL) method proposed in Section 2.3 satisfies the non-decreasing and non-negative constraints and performs better in the weak instrument setting, as well as has the appealing properties of likelihood methods.

2.3. The Maximum Binomial Likelihood Method

The nonparametric MOM uses equations (2.2.14) and (2.2.15) to estimate the CDFs for compliers. Equations (2.2.14) and (2.2.15) identify the CDFs for the compliers at any given point using information on whether outcomes are less than or equal to that point or not. The overall binomial likelihood averages the likelihoods of theses binomial random variables at each point across all observed data points. The advantage of the overall binomial likelihood over the nonparametric MOM is that it ties together the information from equations (2.2.14) and (2.2.15) at each observation point and enforces the constraint that CDFs are nonnegative and nondecreasing. The overall binomial likelihood uses pieces of the true likelihood, but is not equal to the true likelihood because it treats the binomial random variables at each observation point as independent whereas they are actually dependent. In other words, the overall binomial likelihood multiplies together dependent pieces of the true likelihood and is a composite likelihood Lindsay (1988); Varin et al. (2011).

2.3.1. Binomial Likelihood in IV Models

Define $\boldsymbol{\theta} : \mathbb{R} \to [0,1]^4$ such that $\boldsymbol{\theta}(t) = (\theta_{co}^{(0)}(t), \theta_{nt}(t), \theta_{co}^{(1)}(t), \theta_{at}(t))$, where $\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} : \mathbb{R} \to [0,1]$ are the distribution functions of compliers without treatment, never-takers, compliers with treatment, and always-takers, respectively. In this section, we temporarily assume that the proportions of compliance classes, $\boldsymbol{\phi} = (\phi_{co}, \phi_{nt}, \phi_{at})$, are known, and we construct the binomial likelihood of $Y|Z, D$ given the known proportions. In practice, we estimate the proportions from data using Z and D, and use the plug-in estimators obtained in Section 2.2. Given the data, each component of the binomial likelihood is obtained at each data point of Y.

For $u, v \in \{0,1\}$ denote by K_{uv}^a the event $\{Z_a = u, D_a = v\}$. For each pair of outcomes (Y_a, Y_b), define the corresponding component of the binomial likelihood as

$$\begin{aligned} & L_{a,b}(\boldsymbol{\theta}) \\ := & \prod_{u,v\in\{0,1\}} \mathbb{P}(Y_a \le Y_b | Z_a = u, D_a = v)^{\mathbf{1}\{Y_a \le Y_b, Z_a = u, D_a = v\}} \mathbb{P}(Y_a > Y_b | Z_a = u, D_a = v)^{\mathbf{1}\{Y_a > Y_b, Z_a = u, D_a = v\}} \\ = & \prod_{u,v\in\{0,1\}} \theta_{uv}(Y_b)^{\mathbf{1}\{Y_a \le Y_b, Z_a = u, D_a = v\}} (1 - \theta_{uv}(Y_b))^{\mathbf{1}\{Y_a > Y_b, Z_a = u, D_a = v\}} \end{aligned}$$

where the last step uses (2.2.2). We note that this component $L_{a,b}(\boldsymbol{\theta})$ is not symmetric, i.e., $L_{a,b}(\boldsymbol{\theta}) \neq L_{b,a}(\boldsymbol{\theta})$. Then, the binomial likelihood is constructed as the product of all components and is defined as

$$L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}) = \prod_{a=1}^{n} \prod_{b=1}^{n} L_{a,b}(\boldsymbol{\theta}) \tag{2.3.1}$$

Hereafter, for notational brevity, the dependence of the data $\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}$ will be omitted, and the binomial likelihood will be denote by $L(\boldsymbol{\theta})$. Using the identities $F_{01}(t) = F_{at}(t)$, $F_{10}(t) = F_{nt}(t)$, and (2.2.4), (2.2.5), the binomial log-likelihood $\ell(\boldsymbol{\theta})$ can be easily derived as

$$\ell(\boldsymbol{\theta}) := \log L(\boldsymbol{\theta}) = \sum_{a=1}^{n} \sum_{b=1}^{n} \log L_{a,b}(\boldsymbol{\theta}), \tag{2.3.2}$$

From the definition of the component $L_{a,b}(\boldsymbol{\theta})$, we have

$$\begin{aligned} \sum_{a=1}^{n} \log L_{a,b}(\boldsymbol{\theta}) &= \sum_{u,v\in\{0,1\}} n_{uv} \left\{ \mathbb{F}_{uv}(Y_b) \log \theta_{uv}(Y_b) + (1 - \mathbb{F}_{uv}(Y_b)) \log(1 - \theta_{uv}(Y_b)) \right\} \\ &= \sum_{u,v\in\{0,1\}} n_{uv} \cdot J(\mathbb{F}_{uv}(Y_b), \theta_{uv}(Y_b)), \end{aligned}$$

where $\mathbb{F}_{uv}$ is defined in equation (2.2.9) and $J(x,y) := x \log y + (1-x) \log(1-y)$. Therefore,

the binomial log-likelihood can be further simplified as

$$\ell(\boldsymbol{\theta}) = \sum_{b=1}^{n} \sum_{u,v\in\{0,1\}} n_{uv} \cdot J(\mathbb{F}_{uv}(Y_b), \theta_{uv}(Y_b)). \tag{2.3.3}$$

To this end, for $\boldsymbol{\theta} \in \boldsymbol{\vartheta}$, the functional $\mathbb{M}_n(\boldsymbol{\theta})$ is defined as

$$\mathbb{M}_n(\boldsymbol{\theta}) := \frac{1}{n^2}\ell(\boldsymbol{\theta}). \tag{2.3.4}$$

and the maximum binomial likelihood (MBL) estimator of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}_+\in\boldsymbol{\vartheta}} \mathbb{M}_n(\boldsymbol{\theta}). \tag{2.3.5}$$

The MBL estimator is not easily obtained because of the parameter space ϑ_+, especially the non-decreasing condition. We consider an algorithm by combining the EM algorithm and the pool-adjacent-violator algorithm in order to achieve this maximization. We illustrate the details on this algorithm in Appendix A.3.

Now, using (2.3.3), the functional $\mathbb{M}_n(\boldsymbol{\theta})$ can be written as,

$$\mathbb{M}_n(\boldsymbol{\theta}) = \sum_{u,v\in\{0,1\}} T_{uv}^{(n)}(\theta_{uv}), \tag{2.3.6}$$

where

$$T_{uv}^{(n)}(\theta_{uv}) = \frac{1}{n}\sum_{b=1}^{n} \frac{n_{uv}}{n} \cdot J(\mathbb{F}_{uv}(Y_b), \theta_{uv}(Y_b)).$$

Similarly, the *limiting functional* $\mathbb{M}(\boldsymbol{\theta})$ is defined as follows:

$$\mathbb{M}(\boldsymbol{\theta}) = \sum_{u,v\in\{0,1\}} T_{uv}(\theta_{uv}), \tag{2.3.7}$$

where

$$T_{uv}(\theta_{uv}) = \frac{1}{n}\sum_{b=1}^{n} \eta_{00} \cdot J(F_{uv}(Y_b), \theta_{uv}(Y_b)).$$

where η_{uv} is the limit of n_{uv}/n as $n \to \infty$ (see (2.2.7)) and F_{uv} is the population distribution function for the partition $Z = u, D = v$ (see (2.2.4) and (2.2.5)).

2.3.2. Theoretical Results

Fix $0 < \kappa < 1$, and let I_κ be an index set between $\lceil n\kappa \rceil$ and $\lceil n(1-\kappa) \rceil$. That is, $b \in I_\kappa$ implies that b should lie between $\lceil n\kappa \rceil$ and $\lceil n(1-\kappa) \rceil$.

Theorem 2.3.1. *(Consistency) Let* $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\vartheta_+} \mathbb{M}_n(\boldsymbol{\theta}, \boldsymbol{\phi})$ *be the BL estimate,* $\boldsymbol{F}(t) =$

$(F_{co}^{(0)}(t), F_{nt}(t), F_{co}^{(1)}(t), F_{at}(t))'$ *be the population distribution functions. Then*

$$\frac{1}{n}\sum_{b\in I_\kappa} ||\hat{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 = o_P(1). \tag{2.3.8}$$

Theorem 2.3.2. *Under general alternatives, the BL estimate* $\hat{\boldsymbol{\theta}}$ *satisfies*

$$\frac{1}{n}\sum_{b\in I_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{\theta}}(Y_{(b)}) - \hat{\boldsymbol{F}}(Y_{(b)})\}||_2^2 = O_P(n^{-1/2}), \tag{2.3.9}$$

where

$$\hat{\boldsymbol{F}}(t) = \begin{pmatrix} \frac{(\phi_{co}+\phi_{nt})\mathbb{F}_{00}(t) - \phi_{nt}\mathbb{F}_{10}(t)}{\phi_{co}} \\ \mathbb{F}_{10}(t) \\ \frac{(\phi_{co}+\phi_{at})\mathbb{F}_{11}(t) - \phi_{at}\mathbb{F}_{01}(t)}{\phi_{co}} \\ \mathbb{F}_{01}(t) \end{pmatrix}. \tag{2.3.10}$$

Moreover,

$$\frac{1}{n}\sum_{b\in I_\kappa} \left|\left|\sqrt{n}\{\hat{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})\} - \boldsymbol{G}(Y_{(b)})\right|\right|_2^2 = O_P(n^{-1/2}), \tag{2.3.11}$$

where

$$\begin{pmatrix} \frac{1}{\phi_c}\{\sqrt{\frac{\phi_c+\phi_n}{\phi_0}}B_{00}(F_{00}) - \sqrt{\frac{\phi_n}{\phi_1}}B_{10}(F_{10})\} \\ B_{10}(F_{10}) \\ \frac{1}{\phi_c}\{\sqrt{\frac{\phi_c+\phi_a}{\phi_1}}B_{11}(F_{11}) - \sqrt{\frac{\phi_a}{\phi_0}}B_{01}(F_{01})\} \\ B_{01}(F_{01}) \end{pmatrix} \tag{2.3.12}$$

for $B_{00}, B_{01}, B_{10}, B_{11}$ *independent Brownian bridges.*

Theorem 2.3.1 shows that the MBL estimator $\hat{\boldsymbol{\theta}}$ converges to the population distributions $\boldsymbol{F}$. Theorem 2.3.2 shows that the MBL estimator is asymptotically equivalent to the MOM estimator discussed in Section 2.2.2. This is intuitively reasonable because as n goes to infinity, the MOM estimator that is typically not in the parameter space ϑ_+ gradually approaches to the population distribution that already lies in ϑ_+. Since the MBL estimator and the MOM estimator have the same limit distributions, both estimators behave the same way when n is large enough.

However, we emphasize that when n is not large or when an IV is weak (i.e., the proportion of compliers ϕ_{co} is small), the MBL method and the MOM method produce significantly different estimates in finite sample cases. Using simulations, we illustrate this distinction in Appendix A.1 and show that the MBL method produces estimates much close to the population distribution. Figure 1 shows the comparison between the MOM estimate and the MBL estimate when the IV is weak obtained from a simulated dataset. The estimated CDF

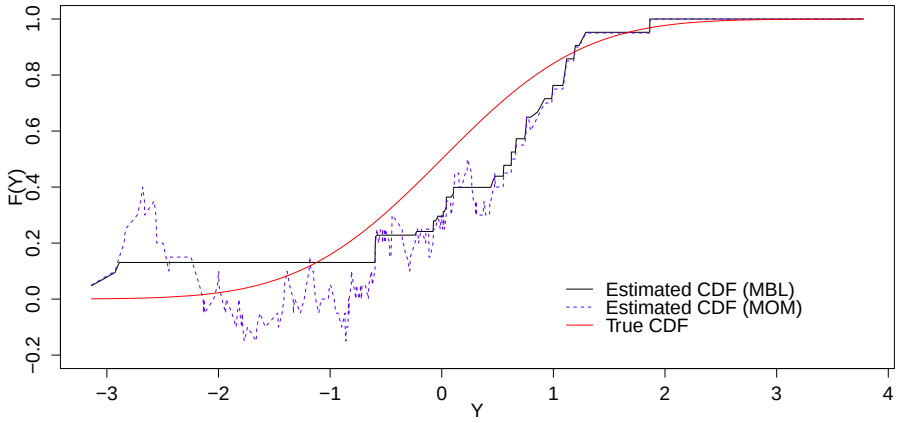

Figure 1: Comparison between estimated CDFs by the MBL method and the MOM method in weak IV setting

from the MBL method, the staircase-like solid curve, is much smoother than the estimated CDF from the MOM method, the dashed curve. It is clear that the MBL estimate is much closer to the population distribution. Also, we can see that the MOM estimate violates the non-decreasing condition for distribution functions.

2.4. Hypothesis Test

2.4.1. Binomial likelihood ratio test

A central question in many studies is, does the treatment have any effect on the distribution of outcomes? Under the IV assumptions, this corresponds to asking, does the treatment have any effect on the distribution of outcomes for compliers? The null hypothesis can be formulated by

$$F_{co}^{(0)}(y) = F_{co}^{(1)}(y) \quad \forall y \in \mathbb{R}. \tag{2.4.1}$$

Since our proposed MBL method can incorporate the constraint $F_{co}^{(0)}(y) = F_{co}^{(1)}(y)$, a new hypothesis test can be created to test (2.4.1). We first review the existing hypothesis test approach and then illustrate our proposed approach.

The test of equality between two distributions is often tested using the Kolmogorov-Smirnov test statistic, which measures the discrepancy between two distributions, i.e.

$$T_{KS} = \sup_y |\theta_{co}^{(0)}(y) - \theta_{co}^{(1)}(y)| \tag{2.4.2}$$

The Kolmogorov-Smirnov (KS) test approach in IV models using the MOM estimates is discussed in Abadie (2002). This test does not bring in the structure of the IV model for

the following reasons. In Section 2.2.2, we obtained the estimates $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$. Since

$$\breve{F}_{co}^{(0)}(y) - \breve{F}_{co}^{(1)}(y) = \frac{\frac{1}{n_1}\sum_{a=1}^{n}\mathbf{1}(Z_a = 1, Y_a \leq y) - \frac{1}{n_0}\sum_{a=1}^{n}\mathbf{1}(Z_a = 0, Y_a \leq y)}{\breve{\phi}_c},$$

the KS test statistic $T_{n,KS}$ of these estimates is

$$T_{n,KS} = \frac{1}{\breve{\phi}_c} \cdot \sup_y \left| \frac{1}{n_1}\sum_{a=1}^{n}\mathbf{1}(Z_a = 1, Y_a \leq y) - \frac{1}{n_0}\sum_{a=1}^{n}\mathbf{1}(Z_a = 0, Y_a \leq y) \right|. \tag{2.4.3}$$

This means that there is no use of the estimates $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$ to conduct the hypothesis test. In other words, the test (2.4.2) proposed by Abadie (2002) is the KS test of whether the distribution of the $Z = 0$ group is the same as the distribution of the $Z = 1$ group. This test is conducted by comparing the empirical distribution function given $Z = 0$ with the empirical distribution function given $Z = 1$, which does not make any use of the structure of the IV model.

In Section 2.3, we estimated all distribution functions, $\hat{\theta}_{co}^{(0)}, \hat{\theta}_{nt}, \hat{\theta}_{co}^{(1)}, \hat{\theta}_{at}$. We extend this approach to estimating the distributions under the null hypothesis $F_{co}^{(0)} = F_{co}^{(1)}$. By enforcing restriction of $\theta_{co}^{(0)} = \theta_{co}^{(1)}$, we can estimate the common distribution of compliers and the distributions of never-takers and always-takers. To be specific, the estimation is achieved by replacing $\theta_{co}^{(0)}$ and $\theta_{co}^{(1)}$ by the common parameter θ_{co} in the binomial likelihood function. Then, the test of equality (2.4.1) can be conducted by a binomial likelihood ratio test which is two times the difference between the overall binomial log-likelihood under the alternative hypothesis and the overall binomial log-likelihood under the null. The binomial likelihood ratio test (BLRT) statistic is formulated by

$$T_{BLRT} = 2 \cdot \left(\max_{\boldsymbol{\theta} \in \vartheta_+} \ell(\boldsymbol{\theta}) - \max_{\boldsymbol{\theta} \in \vartheta_{+,0}} \ell(\boldsymbol{\theta}) \right). \tag{2.4.4}$$

where $\ell(\boldsymbol{\theta})$ is the binomial log-likelihood in IV models defined in Section 2.3.

Theorem 2.4.1. *Let $B_{00}, B_{01}, B_{10}, B_{11}$ be independent Brownian bridges, and*

$$G(t) = \left(\sqrt{\frac{\phi_c + \phi_n}{\phi_0}} B_{00}(F_{00}) - \sqrt{\frac{\phi_n}{\phi_1}} B_{10}(F_{10}) \right) - \left(\sqrt{\frac{\phi_c + \phi_a}{\phi_1}} B_{11}(F_{11}) - \sqrt{\frac{\phi_a}{\phi_0}} B_{01}(F_{01}) \right)$$

Then, for $0 < \kappa < 1$ fixed, under the null hypothesis $H_0 : F_{co}^{(0)} = F_{co}^{(1)}$,

$$T_{BLRT} \to \int_{H^{-1}(\kappa)}^{H^{-1}(1-\kappa)} \frac{G(t)^2}{Var(G(t))} \mathrm{d}H(t), \tag{2.4.5}$$

where $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$.

Table 2: Size and power of test with a significance level $\alpha = 0.05$.

		$N(-\mu,1)$ vs. $N(\mu,1)$			$N(0,1)$ vs. $N(0,\sigma)$	
n	μ	BLRT	KS	σ	BLRT	KS
300	0	0.057	0.054	1	0.048	0.050
300	0.1	0.088	0.067	0.2	0.683	0.604
300	0.2	0.212	0.166	0.4	0.334	0.261
300	0.3	0.340	0.279	0.6	0.135	0.113
300	0.4	0.558	0.479	0.8	0.071	0.076
300	0.5	0.691	0.600	1	0.048	0.050
300	0.6	0.849	0.785	1.2	0.061	0.061
300	0.7	0.911	0.859	1.4	0.117	0.083
300	0.8	0.965	0.926	1.6	0.163	0.093
300	0.9	0.987	0.964	1.8	0.245	0.122
300	1	0.994	0.989	2	0.366	0.140
1000	0.2	0.445	0.383	1.5	0.337	0.169
1000	0.5	0.992	0.976	2	0.932	0.525
2000	0.2	0.719	0.699	2	1.000	0.899

2.4.2. Simulation

To assess the performance of the proposed likelihood ratio test, we compare it to the KS test in Abadie (2002) in a simulation study. The distributions of never-takers and always-takers were fixed as $F_{nt} \sim N(-1,1)$ and $F_{at} \sim N(1,1)$ with the proportions $\boldsymbol{\phi} = (\phi_{co}, \phi_{nt}, \phi_{at}) = (1/3, 1/3, 1/3)$. A sample of size n was drawn and for each sample, the corresponding p-value was (approximately) calculated by using Theorem 2.4.1. To obtain the distribution of the test statistic T_{BLRT} under the null, we replaced the distribution H by the empirical distribution $\mathbb{H}$ (see (2.2.10)) from data. Then, we obtained 10,000 bootstrapped test statistics under the null that provide an approximated p-value for each test statistic obtained from data. We repeated the process of computing a p-value for a simulated dataset 1000 times in order to estimate the power.

Two simulation settings were considered. First, $F_{co}^{(0)}$ and $F_{co}^{(1)}$ have normal distributions with different means, but the same variance. Second, the two distributions of compliers have normal distributions with the same mean, but different variances. Table 2 shows size and power of the BLRT. The first row of the table represents that the true size of the test is approximately equal to the nominal significance level $\alpha = 0.05$. Other rows show that the power of the BLRT is much greater than that of the KS test. The KS test is most sensitive when the distributions differ in a global fashion near the center of the distributions Stephens (1974), which implies that the KS test has good power in the case of an additive treatment effect for a normal distribution. However, even in the favorable case for the KS test, the BLRT is more powerful than the KS test. In the unequal variance case, the BLRT significantly outperforms the KS test.

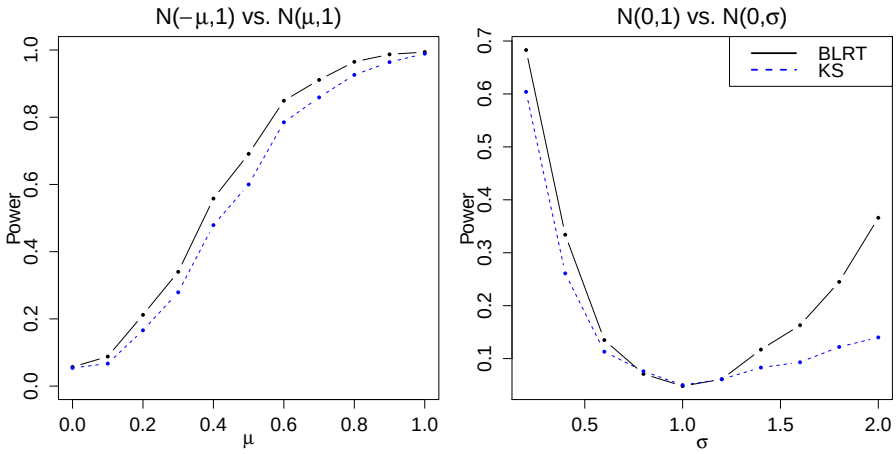

Figure 2: Power of the BLRT test and the KS test. Power is calculated given a significance level $\alpha = 0.05$. Left-panel: the two distributions have different means but the same standard deviation, $F_{co}^{(0)} : N(-\mu, 1)$ versus $F_{co}^{(1)} : N(\mu, 1)$. Right-panel: the two distributions have the same mean but different standard deviations, $F_{co}^{(0)} : N(0, 1)$ versus $F_{co}^{(1)} : N(0, \sigma)$.

Figure 2 visually represents the results of the two simulation settings. The left panel represents the plot of the power versus the size of additive effect of treatment. The BLRT detects the additive treatment effect better than the KS test can. If the treatment effect is large enough, then the powers of the two tests are both close to 1. In the unequal variance case, the gain of the BLRT test over the KS test is even greater as shown in the right-hand panel. In summary, in the simulation setting considered, the BLRT test dominates the KS test.

2.5. ECLS-K 2010-2011: The Effect of SBP Participation on Childhood Obesity

We consider the sample of 2,568 children from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011 (ECLS-K:2011). The children in the ECLS-K:2011 comprise a nationally representative sample selected from both public and private schools attending both full-day and part-day kindergarten in 2010-2011. Using this data, we investigate the effect of participation in the school breakfast program (SBP) on childhood obesity. The treatment is an indicator of SBP participation in Spring 2011 and the outcome is the body mass index (BMI) measured in Fall 2011. There is growing concern that SBP participation contributes to childhood obesity (Ralston et al., 2008; Story et al., 2008). The SBP was initially offered to combat the problem of widespread nutritional deficiencies, but the program more recently has been under the suspicion that it may lead to obesity.

There are two main problems to discuss in order to estimate the effect of participation in the SBP on childhood obesity. First, we need to find a valid IV because participation in the SBP is a non-randomized treatment variable. We follow the suggestion from Jacobson et al. (2001) and consider a binary indicator of living close to school based on the distance from home to school (coded as 1: less than 1/2 miles and 0: 1/2 miles to 5 miles) as an IV. To make plausible causal inference using this IV, we need to verify the validity of the IV. We

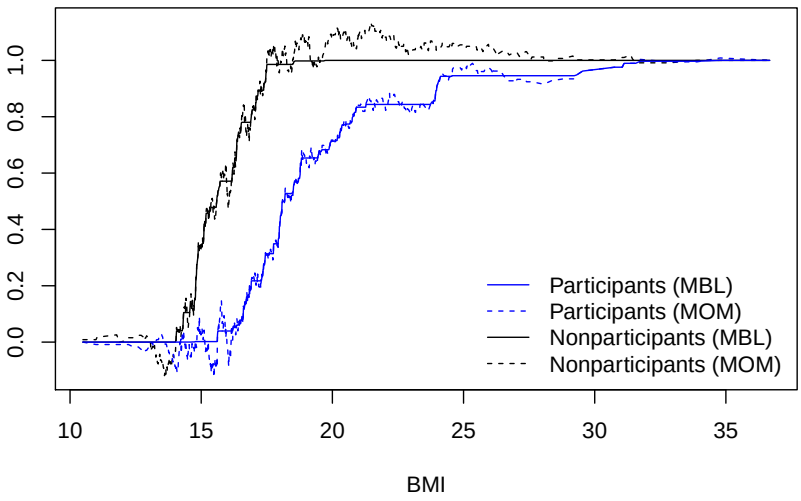

Figure 3: The estimated outcome distributions for compliers given treatment (participants) and no treatment (nonparticipants) obtained from the MBL and MOM methods.

assume that children living close to their schools are more likely to participate in the SBP than children living further away; 48.8% of children who live near their schools participate in the SBP while 41.5% of children who live far from their schools do. This supports the non-zero effect of the IV on the treatment, which satisfies Assumption 1 (3). Also, average BMI measured in Fall 2010, which is before SBP participation, does not significantly differ across children who live far or near their schools (average BMI 16.7 (far) vs. 16.6 (near)). This empirically supports Assumption 1 (5), exclusion restriction assumption because there is no effect of living close to schools on BMI before participating in the SBP. Second, the childhood obesity is not a definite term. The definition of childhood obesity is defined as a BMI at or above the 95th percentile for children of the same age and sex (Barlow and Dietz, 1998). Therefore, the estimation of the average treatment effect on the outcome cannot answer our research question. Instead, we estimate the distributions of the BMI for compliers with treatment and without treatment by using our proposed MBL method.

Figure 3 shows the estimated CDFs of SBP participants and nonparticipants for compliers. We see that the nonparticipants' estimated CDF is almost always above the participants' estimated CDF. The gap between the two CDFs is quite wide. The MBL method improves two features of the MOM method in this example. The MBL estimates for compliers do not violate the nondecreasing and nonnegative conditions. This improvement leads to much smoother CDFs. From satisfying the nondecreasing condition, an additional useful feature of the MBL method is obtained. There is a unique value of estimated earnings corresponding to a specific quantile level. This feature can be useful for those who want to estimate the treatment effect at a certain quantile level using the estimated CDFs. A unique estimate cannot be obtained in the MOM method because of the fluctuation of

Table 3: Tests on distributional effect of the SBP participation on the distribution of the BMI.

	BLRT	KS
p-value	0.013	0.038

the CDF - if there are multiple values that correspond to the same quantile level, then we cannot acquire the corresponding quantile to estimate the causal treatment effect for compliers of that quantile level. For instance, the 95th percentile of the population BMI for children attending kindergarten is estimated as 19.2 (Ogden and Flegal, 2010); this is the cutoff value that defines obesity for these children. The value of the nonparticipants' BMI distribution for compliers at 19.2 is estimated as 0.998 and the corresponding value for the participants is estimated as 0.654. This means that 0.2% of the participants (among compliers) in the SBP have obesity and 34.6% of the nonparticipants (among compliers) have obesity. We emphasize that this inference cannot be made if only the average treatment effect is estimated or the MOM method is used.

Using the hypothesis test approach described in Section 2.4, we can further investigate the distributional treatment effect. We conduct the hypothesis test of no distributional treatment effect, $F_{co}^{(0)} = F_{co}^{(1)}$. Both the new proposed BLRT statistic T_{BLRT} and the KS statistic T_{KS} are considered. Table 3 reports p-values for the two tests of equality. From the p-value approximation scheme described in Section 2.4.2, the p-value of the BLRT is computed by bootstrapping the null distribution of the test statistic. Although the p-values are somewhat different, for both tests, we can reject the null hypothesis that the distributions are equal at a significance level $\alpha = 0.05$. This implies that there is significant evidence that there is an effect of participation in the SBP on the distribution of BMI.

2.6. Summary

We propose the concept of the binomial likelihood to construct nonparametric likelihood by integrating individual likelihoods at all observations. The MBL method produces the estimate that overcomes the limitations of the existing methods. The estimate of the distribution of each compliance classes is used to make genuine nonparametric inference in IV models. Also, our proposed method ensures the properties of non-decreasingness and non-negativeness of distribution functions which have not been achieved for nonparametric IV estimation before. We find that the strength of the MBL method over existing nonparametric methods is particularly pronounced in the weak IV setting. Furthermore, we propose a more powerful test statistic based on the binomial likelihood ratio to test the equality of distributions between treatment and control $F_{co}^{(0)} = F_{co}^{(1)}$.

CHAPTER 3 : Discovering Effect Modification in an Observational Study of Surgical Mortality at Hospitals with Superior Nursing

3.1. Superior Nurse Staffing, Surgical Mortality and Resource Utilization in Medicare

Hospitals vary in the extent and quality of their staffing, technical capabilities and nursing work environments. Does superiority in these areas confer benefits to patients undergoing forms of general surgery that might be performed at most hospitals? To what extent and in what way do these factors affect the cost of surgical care? Are they a life-saving benefit or a pointless and unneeded expense in the case of relatively routine forms of surgery?

A recent study by Silber et al. (2016) sought to answer these questions using Medicare data for Illinois, New York and Texas in 2004-2006. A useful marker for superior staffing is superior nurse staffing, because there is a national voluntary accreditation program to recognize excellent nursing environments, so-called "magnet hospitals"; see Aiken et al. (2000). Additionally, it is relatively easy to use Medicare files to determine the quantity of nurse staffing in the form of the nurse-to-bed ratio. The study compared patient outcomes at 35 magnet hospitals with nurse-to-bed ratios of 1 or more to outcomes for patients at 293 hospitals without magnet designation and with nurse-to-bed ratios less than 1. For brevity, hospitals in the first group are called magnet hospitals and those in the second group are called controls. The question being asked is: How does a patient's choice of hospital, magnet or control, affect the patient's outcomes and medical resource utilization? How consequential is this choice among hospitals and what are its consequences? There is no suggestion, implicit or otherwise, in this question that the nurses are the active ingredient distinguishing magnet and control hospitals, no suggestion that hiring nurses or changing the nurse environment would make control hospitals perform equivalently to magnet hospitals. Magnet designation marks a type of hospital, but does not identify what components are critical in distinguishing that type of hospital. Indeed, the 35 magnet hospitals had many advantages in staffing or technology: 21.5% of magnet hospitals were major teaching hospitals, as opposed to 5.7% of control hospitals; magnet hospitals had more nurses with advanced training, more medical residents per bed, and were somewhat more likely to have a burn unit, and to perform difficult forms of specialist surgery such as coronary bypass surgery and organ transplantation; see Silber et al. (2016), Table 1. Does a patient undergoing perhaps comparatively routine general surgery benefit from all of these capabilities or are they wasted on such a patient?

The distinction in the previous paragraph may be restated as follows. The counter-factual under study is: What would happen to a specific patient if that patient were treated at a hospital having the superior staffing of magnet hospitals when compared to what would happen to this same patient if treated at a control hospital? The counter-factual refers to sending the patient to one hospital or another. What would happen if patients were allocated to existing hospitals in a different way? The counter-factual does not contemplate changing the staffing at any hospital. Beds in hospitals with superior staffing are in limited supply, and it is a matter of considerable public importance that this limited resource be allocated to the patients most likely to benefit from it.

Some patients are in relatively good health and require relatively routine care; perhaps these

patients receive little added benefit from magnet hospitals. Some patients are gravely ill and have poor prospects no matter what care is provided; perhaps these patients also receive little added benefit from magnet hospitals. In contrast, some patients would have poor outcomes with inferior care and would have better outcomes with superior care; perhaps these patients benefit most from treatment in a magnet hospital. Are magnet hospitals more effective for some types of patients than for others? This is the question of effect modification in our title.

Silber et al. (2016) created 25,752 matched pairs of two patients, one undergoing general surgery at a magnet hospital, the other at a control hospital. The two patients in a pair underwent the same surgical procedure as recorded in the 4-digit ICD-9 classification of surgical procedures, a total of 130 types of surgical procedure. Additionally, the matching balanced a total of 172 pretreatment covariates describing the patient's health prior to surgery; see Silber et al. (2016) Table 2. Overall, Silber et al. found significantly lower mortality at magnet hospitals than at control hospitals (4.8% versus 5.8%, McNemar P-value < 0.001), substantially lower use of the intensive care unit or ICU (32.9% versus 42.9%) and slightly shorter length of stay; see Silber et al. (2016), Table 3 where costs and Medicare payments are also evaluated. Magnet hospitals had lower mortality rates while making less use of an expensive resource, the ICU.

In one analysis, Silber et al. (2016) grouped matched pairs based on an estimated probability of death that was controlled by the matching algorithm. The lowest risk patients appeared to benefit least from magnet hospitals. In contrast, the fourth quintile of risk — a high, but not the highest quintile of risk — had both lower mortality and lower cost in magnet hospitals, whereas the highest risk quintile had lower mortality but higher cost at magnet hospitals. In brief, Silber et al. (2016) found evidence of effect modification.

Patients with very different medical problems may have similar probabilities of death. It is interesting that the effect of magnet hospitals appears to vary with patient risk, but it would be more interesting still to unpack patient risk into its clinical constituents, and to understand how the effect varies with these constituents. Clinicians do not think of patients in terms of their probability of death, but rather in terms of their specific health problems that are aggregated by the probability of death. In that sense, the examination of effect modification in Silber et al. (2016) is too limited to guide practice.

The current paper uses a recently proposed exploratory technique to unpack effect modification, combined with a confirmatory technique that examines the sensitivity of these conclusions to unmeasured biases. Is the ostensible effect larger, more stable or more insensitive to unmeasured bias for certain surgical procedure clusters or certain categories of patients defined by other health problems?

3.2. Review of Effect Modification in Observational Studies

3.2.1. Notation for causal effects, nonrandom treatment assignment, sensitivity analysis

In observational studies, it is known that certain patterns of treatment effects are more resistant than others to being explained away as the consequence of unmeasured biases

in treatment assignment; see, for instance, Rosenbaum (2004), Zubizarreta et al. (2013), Stuart and Hanna (2013).

Effect modification occurs when the size of a treatment effect or its stability varies with the level of a pretreatment covariate, the effect modifier. Effect modification affects the sensitivity of ostensible treatment effects to unmeasured biases. Other things being equal, larger or more stable treatment effects are insensitive to larger unmeasured biases; see Rosenbaum (2004), Rosenbaum (2005). As a consequence, discovering effect modification when it is present is an important aspect of appraising the evidence that distinguishes treatment effects from potential unmeasured biases, a concern in every observational study. In particular, Hsu et al. (2013, 2015) discuss sensitivity analysis in observational studies with potential effect modification, and §3.2 is a concise summary. Chesher (1984), Crump et al. (2008), Lehrer et al. (2016), Lu and White (2015), Wager and Athey (2015), Athey and Imbens (2016), Ding et al. (2015), discuss effect modification from a different perspective, placing less emphasis on its role in confirmatory analyses that distinguish treatment effects from unmeasured biases in observational studies.

There are I matched pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated with $Z_{ij} = 1$, the other control with $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. Subjects are matched for an observed covariate $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i$, say, for each i, but may differ in terms of a covariate u_{ij} that was not measured. Each subject has two potential responses, r_{Tij} if treated, r_{Cij} if control, exhibiting response $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, so the effect caused by the treatment, $r_{Tij} - r_{Cij}$ is not seen from any subject; see Neyman (1923, 1990) and Rubin (1974). Fisher (1935) null hypothesis H_0 of no treatment effect asserts $r_{Tij} = r_{Cij}$ for all i, j. Simple algebra shows that the treated-minus-control pair difference in observed responses is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$ which equals $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) = \pm\,(r_{Ci1} - r_{Ci2})$ if Fisher's hypothesis H_0 is true. Write $\mathcal{F} = \{(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij},\, u_{ij}),\, i = 1, \ldots, I,\, j = 1, 2\}$ for the potential responses and covariates, and write $\mathcal{Z}$ for the event that $Z_{i1} + Z_{i2} = 1$ for each i.

In a randomized experiment, $Z_{i1} = 1 - Z_{i2}$ is determined by I independent flips of a fair coin, so $\pi_i = \Pr(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ for each i, and this becomes the basis for randomization inferences, for instance for tests of Fisher's null hypothesis or for confidence intervals or point estimates formed by inverting hypothesis tests. A randomization inference derives the null distribution given $(\mathcal{F}, \mathcal{Z})$ of a test statistic as its permutation distribution using the fact that the 2^I possible values of $\mathbf{Z} = (Z_{i1}, Z_{i2}, \ldots, Z_{I2})$ each have probability 2^{-I} in a randomized paired experiment; see Fisher (1935), Lehmann and Romano (2005), or Rosenbaum (2002b). A simple model for sensitivity analysis in observational studies says that treatment assignments in distinct pairs are independent but bias due to nonrandom treatment assignment may result in π_i that deviate from $\frac{1}{2}$ to the extent that $1/(1+\Gamma) \leq \pi_i \leq \Gamma/(1+\Gamma)$ for $\Gamma \geq 1$, and the range of possible inferences is reported for various values of Γ to display the magnitude of bias that would need to be present to materially alter the study's conclusion; see, for instance, Rosenbaum (2002a) for the case of matched binary responses, as in the current paper. For instance, a sensitivity analysis may report the range of possible P-values or point estimates that are consistent with the data and a bias of at most Γ for several values of Γ.

For various approaches to sensitivity analysis in observational studies, see Cornfield et al. (1959), Gastwirth (1992), Gilbert et al. (2003), Egleston et al. (2009), Hosman et al. (2010), Liu et al. (2013). For some discussion of software in R, see Rosenbaum (2015) and Rosenbaum and Small (2016).

3.2.2. Three Strategies Examining Effect Modification

There is effect modification if the magnitude of the effect, $r_{Tij} - r_{Cij}$, varies systematically with $\mathbf{x}_i$. We partition the space of values of $\mathbf{x}_i$ into subsets and are concerned with effects that differ in magnitude or stability between subsets. Let $\mathcal{G}$ be a subset of the values of $\mathbf{x}$, and define the null hypothesis $H_{\mathcal{G}}$ to be Fisher's null hypothesis for individual j in set i with $\mathbf{x}_{ij} \in \mathcal{G}$, so $H_{\mathcal{G}}$ asserts that $r_{Tij} = r_{Cij}$ for all ij with $\mathbf{x}_{ij} = \mathbf{x}_i \in \mathcal{G}$. Let $\mathfrak{g} = (\mathcal{G}_1, \ldots \mathcal{G}_G)$ be a mutually exclusive and exhaustive partition of values of $\mathbf{x}_{ij} = \mathbf{x}_i$, so each pair i has an $\mathbf{x}_i$ contained in exactly one $\mathcal{G}_g$. A simple form of effect modification occurs if $H_{\mathcal{G}_g}$ is true for some g but not for other g. Write I_g for the number pairs with $\mathbf{x}_i \in \mathcal{G}_g$, so $I = \sum_{g=1}^{G} I_g$.

There are three strategies for defining the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, two of which are practically useful but technically straightforward, the third having interesting technical aspects that we illustrate using the Medicare example. One useful strategy defines the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, a priori, without reference to data. For example, on the basis of clinical judgement, one might believe certain surgical procedures are more challenging or hazardous than others, and therefore divide the exactly matched procedures into a few groups based on clinical judgement alone. Alternatively, clinical judgement might separate patients with severe chronic conditions unrelated to the current surgery, such as congestive heart failure.

A second strategy uses an external source of data to define the groups. In particular, Silber et al. (2016) fit a logit model to an external data source, predicting mortality from covariates, $\mathbf{x}_{ij}$, then formed five groups $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_5)$ based on this predicted risk for a given $\mathbf{x}$. This approach made no use of the mortality experience of the patients in the current study in defining the groups. A variant of the second strategy is to split one data set at random into two parts, create the groups using the first part, then analyze only the second part with these, again, externally determined groups.

In both of the first two strategies, the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, were determined by events external to the outcomes reported study. The second strategy makes explicit use of an external source of data, while the first strategy uses judgement that is presumably informed historically by various external sources of data. The key element in both strategies is that the groups were fixed before examining outcomes in the current study, and in that sense are unremarkable as groups, requiring no special handling because of their origin. With a priori groups, we could use any of a variety of methods to test the G hypotheses $H_{\mathcal{G}_g}$ in such a way as to strongly control the family-wise error rate at α, meaning that the chance of falsely rejecting at least one true $H_{\mathcal{G}_g}$ is at most α no matter which hypotheses are true and which are false.

The third strategy that we illustrate here creates the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, by exploratory techniques using all of the current data, and then goes on to perform an analysis

of the same data as if the groups had been determined a priori. The third strategy is designed so that it controls the family-wise error rate in a sensitivity analysis despite the data-dependent generation of G particular groups from among the infinitely many ways of splitting the space of values of the observed covariates $\mathbf{x}$. This strategy is discussed in detail in Hsu et al. (2015) and it entails certain restrictions on the way the groups are constructed.

A simple version of the strategy regresses $|Y_i| = |(Z_{i1} - Z_{i2})(R_{i1} - R_{i2})|$ on $\mathbf{x}_i$ using a form of regression that yields groups, such as CART. For discussion of CART, see Breiman et al. (1984) and Zhang and Singer (2010). Note that the unsigned $|Y_i|$ not the signed Y_i are used; that is, the regression does not know who is treated and who is control. The leaves of a CART tree become the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$. The signs of the Y_i are then "remembered," in an analysis that views the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, as fixed, so it resembles analyses that would have been appropriate with an a priori grouping of the type created by the first two strategies.

It is important to understand what is at issue in the third strategy; see Hsu et al. (2015) for a precise and general technical discussion. Briefly if obscurely, the groups, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$, and hence the hypotheses, $H_{\mathcal{G}_g}$, are not stable. If the observed data had been slightly different, the CART tree would have been different, and we would be testing different hypotheses. What does it mean to speak about the probability of falsely rejecting $H_{\mathcal{G}_g}$ if most data sets would not lead us to test $H_{\mathcal{G}_g}$?

Consider the simplest case, a paired randomized experiment. If Fisher's null hypothesis of no effect of any kind were true, then $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \pm(r_{Ci1} - r_{Ci2})$ and, given $(\mathcal{F}, \mathcal{Z})$, different random assignments Z_{ij} always yield $|Y_i| = |r_{Ci1} - r_{Ci2}|$, so all 2^I random assignments produce the same CART tree and the same $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$. In other words, under H_0, the CART tree and hence $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$ is a function of $(\mathcal{F}, \mathcal{Z})$ and not of $\mathbf{Z}$. Therefore, under H_0, the 2^{I_g} possible treatment assignments for the I_g pairs with $\mathbf{x}_i \in \mathcal{G}_g$ each have probability 2^{-I_g}, resulting in conventional permutation tests within each of the G groups, tests that are conditionally independent given $(\mathcal{F}, \mathcal{Z})$ under H_0. The problem occurs because we are interested in testing not just H_0, but also individual $H_{\mathcal{G}_g}$ when H_0 is false because some individuals are affected by the treatment. If H_0 is false, different random assignments $\mathbf{Z}$ yield different $|Y_i|$, hence different CART trees and different hypotheses, $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$. With a bit of care, it is possible to demonstrate two useful facts. First, if $r_{Tij} - r_{Cij} = 0$ for all ij with $\mathbf{x}_i \in \mathcal{G}_g$, then the conditional distribution given $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$ and $(\mathcal{F}, \mathcal{Z})$ of the corresponding Z_{ij} with $\mathbf{x}_i \in \mathcal{G}_g$ is its usual randomization distribution. In that sense, the instability of the tree over repeated randomizations has not distorted this conditional distribution of treatment assignments in groups with no treatment effect. Second, if a method is applied to test the $H_{\mathcal{G}_g}$ that would strongly control the family-wise error rate at α with a priori fixed groups, then conditionally given $\mathfrak{g} = (\mathcal{G}_1, \ldots, \mathcal{G}_G)$ and $(\mathcal{F}, \mathcal{Z})$, the method will reject at least one null group with probability at most α. These two facts are extended to include sensitivity analyses in observational studies and are proved as Propositions 1 and 2 of Hsu et al. (2015). That paper also presents some reasons to hope that subsets of $\mathbf{x}_i$ that systematically predict $|Y_i|$ may identify groups in which either the magnitude of $r_{Tij} - r_{Cij}$ or its stability varies

with $\mathbf{x}_i$.

In the current paper, we present a practical application of this third strategy.

3.3. Discovering and Using Effect Modification in the Magnet Hospital Study

3.3.1. Forming Groups of Pairs for Consideration as Possible Effect Modifiers

The analyses here first broke and then re-paired the pairs in Silber et al. (2016) so that: (i) as in Silber et al., every pair was exactly matched for the 130 four-digit ICD-9 surgical procedure codes, (ii) the maximum number of pairs were exactly matched for an indicator of age greater than 75, congestive heart failure (CHF), emergency admission or not, and chronic obstructive pulmonary disease (COPD). Because identically the same people were paired differently, the balancing properties of the new pairs are exactly the same as reported by Silber et al. (2016, Table 2), because balancing properties refer to marginal distributions of covariates and do not depend upon who is paired with whom.

Using `rpart` in `R`, the CART tree was built using the 22,622 pairs that were exactly matched in the sense described in the previous paragraph, regressing $|Y_i|$ on $\mathbf{x}_i$, where Y_i records the difference in binary indicators of mortality. So, the tree is essentially trying to locate pairs discordant for mortality, $|Y_i| = 1$, on the basis of exactly matched covariates. Here, a pair is discordant if exactly one patient in the pair died within 30-days. CART was not offered all 130 exactly matched surgical procedure codes, but rather 26 mutually exclusive clusters of the 130 surgical procedures, as listed in Table 4, plus the binary covariates age>75, CHF, emergency admission, and COPD. The resulting tree is depicted in Figure 4. A few procedure clusters — e.g., liver procedures — are diverse, perhaps meriting further subdivision that we do not consider here.

We began with 25,752 matched pairs. As described above, only the 22,622 pairs that were exactly matched for five potential effect modifiers were used to build the tree in Figure 4. Ultimately CART used three of the five covariates and ignored the remaining two covariates, namely 'age>75' and COPD. To use the classification in Figure 4, we need pairs that are exactly matched for three covariates, not for five covariates. Can we recover some of the pairs that we did not use because they were not matched for five covariates? To recover omitted pairs, we followed the tactic in Hsu et al. (2015). Specifically, we re-paired as many of the pairs that were not used to build the tree to be exact for the 130 procedures plus CHF and emergency admission, adding these additional 1,093 pairs to the groups in Figure 4, making 23,715 pairs in total, or 95% of the original study. All analyses that follow refer to these 23,715 pairs.

Consider the tree in Figure 4, starting from its root at the top of the figure. The tree split the population into two groups, patients without congestive heart failure (CHF) and patients with CHF, a serious comorbid condition. It then split this divided population by grouping the 26 surgical procedure clusters. There are, of course, many way to group 26 procedure clusters; for instance, there are $2^{26} - 1 = 67,108,863$ ways to split them into two groups. There are four groups of procedures, two for patients with CHF and two for patients without CHF. Table 4 displays CART's grouping of the 26 procedure clusters into

proc1, proc2, proc3 and proc4. In Figure 4, CART further divided proc2 into two subsets of patients, those admitted as emergencies and the remaining nonemergent patients. In Table 4, notice that proc1 and proc3 overlap extensively, as do proc2 and proc4. To the clinical eye, with a few raised eyebrows, the procedures in proc2 and proc4 look riskier or more complex than those in proc1 and proc3. Groups proc1 and proc3 are very similar but not identical, and groups proc2 and proc4 are very similar but not identical. For instance, appendectomy is grouped with the less risky procedures in proc1 if the patient does not have CHF, but it was grouped with the more risky procedures in proc4 for a patient with CHF; however, it is unclear whether that switch is a profound insight or a hiccup.

The CART tree was built by predicting $|Y_i|$ from x_i. In contrast, hypothesis testing will use the signed value of Y_i.

3.3.2. Informal Examination of Outcomes

In §3.3.3, an analysis of mortality is carried out as proposed in Hsu et al. (2015). This analysis is easier to understand if we take a quick look first. The upper part of Table 5 describes mortality informally. The first three numeric rows of Table 5 describe information that CART could use in building the tree, namely the number of pairs, the number of discordant pairs, and the proportion of discordant pairs. In Table 5, 43% = 10127/23715 of pairs are in the group 1, that is, patients without CHF undergoing less risky procedures. Expressed differently, group 1 has the most pairs and the fewest discordant pairs of the five groups. As one might expect given the information that CART was permitted to use, the proportion of discordant pairs varies markedly among the groups CART built.

The next three numeric rows of Table 5 display outcomes by treatment group, making use of Y_i and not just $|Y_i|$. The mortality rates for magnet and control groups are given, as is the odds ratio computed from discordant pairs; see Cox (1970). All of the odds ratios are greater than or equal to 1, suggesting higher mortality at control hospitals. The largest odds ratio is in group 2, 1.53, while the largest difference in mortality rates is in group 5, 18.6%-16.5% = 2.1%. The odds ratio closest to 1 is in group 3, the group most similar to group 2 except for admission through the emergency room.

3.3.3. Structured Analysis of Outcomes in Discovered Groups

The structured analysis in Hsu et al. (2015) starts by computing randomization tests and upper sensitivity bounds on P-values for each of the five groups separately. In Table 5, these are based on a test of the McNemar type, essentially binomial calculations using discordant pairs; see Cox (1970) for discussion of paired binary data, and see Rosenbaum (2002b), §4.3.2 for the sensitivity analysis. In the bottom part of Table 5 are upper bounds on one-sided P-values testing no treatment effect in a group in the presence of a bias in treatment assignment of at most Γ. Also given in Table 5 are the odds ratios from discordant pairs associated with McNemar's test.

The final column in the bottom of Table 5 gives the P-value for the truncated product of P-values as proposed by Zaykin et al. (2002). The truncated product generalizes Fisher's method for combining independent P-values: the test statistic is the product of those P-

values that are smaller than a threshold, τ, where $\tau = 0.1$ in Table 5. Zaykin et al. (2002) determined the null distribution of the truncated product statistic. Hsu et al. (2013) show that the same null distribution may be used to combine upper bounds on P-values in a sensitivity analysis for a tree like Figure 4, and that it often has superior power in this context compared to Fisher's product of all P-values, essentially because sensitivity analyses promise P-values that are stochastically larger than uniform for a given Γ. Truncation eliminates some very large upper bounds on P-values.

Hsu et al. (2015) combine the truncated product statistic with the closed testing procedure of Marcus et al. (1976) to strongly control the family-wise error rate at α in a sensitivity analysis with a bias of at most Γ. Given G hypotheses, $H_{\mathcal{G}_g}$, $g = 1, \ldots, G$, asserting no effect in each of G groups, closed testing begins by defining $2^G - 1$ intersection hypotheses, $H_{\mathcal{L}}$, where $\mathcal{L} \subseteq \{1, \ldots, G\}$ is a nonempty set, and $H_{\mathcal{L}}$ asserts that $H_{\mathcal{G}_\ell}$ is true for every $\ell \in \mathcal{L}$. Closed testing rejects $H_{\mathcal{L}}$ if and only if the P-value testing $H_{\mathcal{K}}$ is $\leq \alpha$ for every $\mathcal{K} \supseteq \mathcal{L}$. The P-value testing $H_{\mathcal{K}}$ is based on the truncated product of P-values for $H_{\mathcal{G}_k}$ for $k \in \mathcal{K}$.

The P-value in the final column of Table 5 tests Fisher's hypothesis H_0, or $H_{\mathcal{L}}$ with $\mathcal{L} = \{1, 2, 3, 4, 5\}$. For $\Gamma = 1$, this test combines five McNemar tests using the truncated product, and in the absence of bias, the hypothesis H_0 is rejected with a one-sided P-value of 2.7×10^{-6}. To complete closed testing of subhypotheses, one performs $2^5 - 1 = 31$ tests of intersection hypotheses. Hypothesis $H_{\{3,4\}}$ has a P-value using the truncated product of 0.080, so neither $H_{\mathcal{G}_3}$ nor $H_{\mathcal{G}_4}$ is rejected at the 0.05 level by closed testing, but $H_{\mathcal{G}_1}$, $H_{\mathcal{G}_2}$ and $H_{\mathcal{G}_5}$ are rejected. In short, in the absence of bias, $\Gamma = 1$, the hypothesis of no effect is rejected in groups 1, 2, and 5.

At $\Gamma = 1.05$, Fisher's hypothesis of no effect at all is rejected at the 9.0×10^{-5} level, and closed testing rejects both $H_{\mathcal{G}_1}$ and $H_{\mathcal{G}_2}$ at the 0.05 level. At $\Gamma = 1.10$, Fisher's hypothesis H_0 of no effect is rejected at the 0.012 level, but only $H_{\mathcal{G}_2}$ is rejected at the 0.05 level. At $\Gamma = 1.17$, Fisher's hypothesis H_0 of no effect is rejected at the 0.044 level, no individual subgroup hypothesis is rejected at the 0.05 level, but $H_{\{1,2\}}$ is rejected at the 0.05 level. At $\Gamma = 1.18$, no hypothesis is rejected at the 0.05 level.

A bias of $\Gamma = 1.17$ corresponds with an unobserved covariate that doubles the odds of having surgery at a control hospital and increases the odds of death by more than 60%. That is, stated technically, $\Gamma = 1.17$ amplifies to $(\Lambda, \Delta) = (2.0, 1.61)$; see Rosenbaum and Silber (2009a). McNemar's test applied to all 23,715 pairs yields a P-value bound of 0.063 at $\Gamma = 1.15$, so this overall test is slightly more sensitive to unmeasured biases and provides no information about subgroups.

What range of possible unmeasured biases, measured by Γ, should be explored? We do not know and cannot know how much bias is actually present in an observational study. However, in a straightforward way, we can and should determine the quantity of bias that would need to be present to alter the study's conclusions, for instance the bias that might lead to acceptance of a null hypothesis rejected in a conventional analysis that assumed no bias, $\Gamma = 1$. The degree of sensitivity to bias is a fact in the data brought to light by an appropriate analysis.

3.3.4. Use of the Intensive Care Unit (ICU)

In Table 5, magnet hospitals exhibited lower mortality than control hospitals for ostensibly similar patients undergoing the same surgical procedure, that is, magnet hospitals exhibited better quality. Does better quality cost more? For resources that are allocated by a market mechanism — say, restaurants or hotels — we expect better quality to cost more, but market forces play little role in Medicare payments. In the absence of market forces, it is an open question whether better quality costs more. Silber et al. (2016) examine this issue in several ways, but Table 6 restricts attention to the consumption of a particularly expensive resource, namely use of the intensive care unit or ICU. In a hospital with inadequate nursing staff, a patient may be placed in the ICU to ensure that the patient is monitored, while in a hospital with superior nursing this same patient might remain in a conventional hospital room. This is one mechanism by which better quality — lower mortality rates — might cost less, not more.

Is the lower mortality in magnet hospitals associated with greater use of the ICU? Apparently not. Overall and in all five groups in Figure 4, the use of the ICU in Table 6 is lower at magnet hospitals than at control hospitals. The odds ratio is largest in group 2, but it is not small in any group. In various other ways also, Silber et al. (2016) found that costs were lower at hospitals with superior nursing, despite lower mortality rates.

The closed testing procedure applied to the sensitivity analysis in the bottom part of Table 6 rejects the null hypothesis of no effect on ICU utilization in all five groups providing the bias in treatment assignment is at most $\Gamma = 1.5$. Using the method in Rosenbaum and Silber (2009a), a bias of $\Gamma = 1.5$ corresponds with an unobserved covariate that increases the odds of surgery at a control hospital by a factor of 4 and increases the odds of going to the ICU by a factor of 2. Closed testing rejects no effect only in group 2 for $1.6 \leq \Gamma \leq 1.8$, and cannot reject even Fisher's H_0 for $\Gamma = 1.9$. Detailed results for group 2 are given in Table 7.

To emphasize a point emphasized in §3.1, Tables 5, 6 and 7 concern the effect of going to a magnet hospital rather than a control hospital for surgery, but they do not show the specific role of nurses in this effect. It is entirely plausible that superior nurse staffing would permit more patients to stay out of the ICU, but nothing in the data speaks to this directly. The main difference between the ICU and the floor of the hospital is the higher density, often higher quality, of the nurse staffing in the ICU. A hospital with a higher nurse-to-bed ratio and superior nurse staffing may be able to care for a seriously ill patient on the hospital floor, where some other hospital would be forced to send the same patient to the ICU.

3.3.5. Other Analyses and Options for Analysis

The tree in Figure 4 was built for mortality, but was used also for ICU use. In an additional analysis, we applied CART to each leaf of Figure 4 to predict unsigned discordance for ICU use. The two interesting aspects of this analysis were: (i) subgroup 2 in Figure 4 was not further divided; (ii) subgroup 5 in Figure 4 was further divided, with more evidence of an effect on ICU use among patients in this subgroup who were not admitted through the emergency room, a pattern analogous to subgroups 2 and 3. An interesting feature of this

type of analysis is that it makes mortality the primary endpoint, as it would be in most surgical studies, so only mortality determines the initial tree for the mortality analysis, but it permits the secondary outcome of ICU use to affect a secondary tree.

We let CART build the groups. Any analysis that used only $|Y_i|$ and $\mathbf{x}_i$ could be used to build the groups. In saying this, we mean that the strong control of the family-wise error rate in Hsu et al. (2015) would not be affected by revisions to the tree that used only $|Y_i|$ and $\mathbf{x}_i$. Indeed, a surgeon who did not look at Y_i could look at Figure 4 and Table 4 and decide to regroup some of the procedure groups. Perhaps the surgeon would view some of CART's decisions as clinically unwise and would change them, or perhaps the surgeon would prefer that proc1 and proc3 be identical, and that proc2 and proc4 be identical. Indeed, the surgeon might suggest fitting the tree again, using only $|Y_i|$ and $\mathbf{x}_i$, but subdividing some procedure clusters, say liver procedures, that seem too broad to be clinically meaningful. What is critical is that the groups are formed using $|Y_i|$ and $\mathbf{x}_i$ without using the sign of Y_i.

3.4. Summary and Discussion: Confirmatory Analyses that Discover Larger Effects by Exploratory Methods

3.4.1. Summary: It is Important to Notice Subgroups with Larger Treatment Effects in Observational Studies

In an observational study of treatment effects, there is invariably concern that an ostensible treatment effect is not actually an effect caused by the treatment, but rather some unmeasured bias distinguishing treated and control groups. Larger or more stable treatment effects are more insensitive to such concerns than smaller or more erratic effects; that is, larger biases measured by Γ would need to be present to explain a large and stable treatment effect. These considerations motivate an interest in effect modification in observational studies. Perhaps the treatment effect is larger or more stable in certain subgroups defined by observed covariates. If so, the ostensible treatment effect in such subgroups is likely to be insensitive to larger unmeasured biases, therefore more credible, and additionally, a larger or more stable effect is likely to be more important clinically.

The magnet hospitals had lower mortality overall, and lower or equivalent mortality in each of the five groups. However, the superior staffing of magnet hospitals was least sensitive to unmeasured bias in our group 2, consisting of patients undergoing relatively serious forms of surgery in the absence of other life-threatening conditions, such as congestive heart failure or an emergency admission leading to surgery. Moreover, not only were mortality rates lower in magnet hospitals for these patients (2.5% rather than 3.5%), but additionally the magnet hospitals cared for these patients with greatly reduced use of an expensive resource, namely the intensive care unit (ICU rate of 28.9% rather than 43.3%). Determining the cost of hospital care for Medicare patients is not straightforward, so Silber et al. (2016) contrasted several formulas to appraise the cost of magnet hospitals. In all of these formulas, use of the ICU plays a substantial part, as does the length of stay in the hospital. Regardless of which formula was used, magnet hospitals appear to produce lower mortality either at no additional cost or with a cost savings.

A plausible interpretation of Figure 4, Table 4 and Table 5 is that: (i) patients in groups 2, 4 and 5 should be directed to magnet hospitals, a limited resource; (ii) the large number of comparatively healthy patients requiring simpler surgical procedures may go to non-magnet hospitals if space in a magnet hospital is unavailable, (iii) patients in group 3 requiring emergency surgery should go to the nearest hospital.

3.4.2. Exploration and Confirmation Using Regression Trees

The CART method of Breiman et al. (1984), as originally proposed, did not lend itself to conventional inference, such as hypothesis testing, much less to simultaneous inference for the groups it produced. In contrast, Hsu et al. (2015) proposed a way to use CART, or similar methods, combining exploratory construction of groups together with a confirmatory sensitivity analysis that controls the family-wise error rate in the constructed groups. All of the data are used to build the tree and all of the data are used in confirmatory analyses. This is important in Table 5 because the number of pairs discordant for mortality is not large in some groups — happily, most people survive surgery — so sample splitting to build the groups would leave less data for confirmatory analyses. The double use of all of the data works by having CART predict $|Y_i|$ from x_i without knowing who is treated and who is control, then using the signed Y_i in confirmatory analyses with CART's groups. CART trees can be unstable, so the tree should be regarded as an interesting partition of the data, not a search for a "true" partition. The formal hypothesis tests are conditional inferences given CART's partition: they correctly use, but do not endorse, the partition.

Will a tree built from $|Y_i|$ be useful in the study of effect modification? It is straightforward to construct theoretical examples in which an analysis of $|Y_i|$ would miss effect modification that an analysis of Y_i might find. Obviously, a tree built from all of the Y_i is preferable, but this would preclude a confirmatory analysis using the same data. As noted by Hsu et al. (2015), a result of Jogdeo (1977), Theorem 2.2 provides some encouragement. A simple version of this result says: if $Y_i = \mu_i + \epsilon_i$, $\mu_i \geq 0$, $i = 1, \ldots, I$, where the errors ϵ_i are independent and identically distributed with a unimodal distribution symmetric about zero, then $|Y_i|$ is stochastically larger than $|Y_j|$ whenever $\mu_i > \mu_j$. Under this simple model, trees that form groups from the level of $|Y_i|$ have some hope of finding groups heterogeneous in μ_i. True, if the ϵ_i are not identically distributed, if the dispersion of ϵ_i varies with i, then the groups may be affected by both level and dispersion; however, sensitivity to unmeasured bias is also affected by both the level and dispersion of the treatment effects, so groups reflecting unequal dispersion are interesting also. For additional encouragement, see also the simulation results in Hsu et al. (2015).

3.4.3. Other Applications

We discussed in detail a clinical application that extends results in Silber et al. (2016). However, the proposed method is also applicable outside clinical medicine. Hsu et al. (2013) presented an example from public health, in which treatments intended to prevent malaria in Nigeria were much more effective for children than for adults, so the conclusions were much more insensitive to unmeasured bias for children than for adults. Hsu et al. (2015) presented an example from labour economics concerning the effect of the 2010 Chilean earthquake on work income. Does a major disaster create employment or interfere with it?

They found that the earthquake had its largest and least sensitive effects on men who had no work income prior to the earthquake: these men were less likely to find jobs and secure work income than similar men who were unaffected by the earthquake.

3.4.4. Alternative Methods

As noted in §3.2.2, there are at least three basic approaches to confirmatory sensitivity analyses for effect modification. One approach starts with a priori groups, or, what amounts to the same thing, groups built from one or more external data sets. Essentially this approach was used in Silber et al. (2016) for these data. The five groups were defined by quintiles of risk-of-death as estimated using a model fit to another set of data. That analysis was enlightening, but the plausible interpretation at the end of §3.4.1 makes useful distinctions that risk quintiles do not make.

Another approach is to: (i) split the data into two parts at random, (ii) form patient groups from Y_i rather than $|Y_i|$ using the first part of the data, (iii) discard the first part, (iv) perform a confirmatory analysis on the second part using the patient groups formed from the first part. This approach is attractive when I is very large. For some indirectly related theory, see Heller et al. (2009). Presumably, if we had twice as many pairs as we actually had, $I \to 2I$, if we split the data in half as just described, then the resulting analysis would be uniformly better than the analysis we did with half as much data, because: (i) the tree would be better having been built from Y_i instead of $|Y_i|$, but (ii) the confirmatory analysis would have the same quantity of data as our confirmatory analysis. Silber et al. (2016) used data from New York, Illinois and Texas primarily because purchasing Medicare data is expensive. There are, however, 47 more states where these came from.

Table 4: Grouping of procedure clusters, with and without congestive heart failure (CHF).

	Procedure Cluster	No CHF proc1	CHF proc3	No CHF proc2	CHF proc4
1	Adrenal procedures	x	x		
2	Appendectomy	x			x
3	Bowel anastamoses			x	x
4	Bowel procedures, other			x	x
5	Breast procedures	x	x		
6	Esophageal procedures		x	x	
7	Femoral hernia procedures	x	x		
8	Gallbladder procedures	x	x		
9	Incisional and abdominal hernias	x	x		
10	Inguinal hernia procedures	x	x		
11	Large bowel resection			x	x
12	Liver procedures	x			x
13	Lysis of adhesions			x	x
14	Ostomy procedures			x	x
15	Pancreatic procedures		x	x	
16	Parathyroidectomy	x	x		
17	PD access procedure			x	x
18	Rectal procedures	x	x		
19	Repair of vaginal fistulas	x	x		
20	Small bowel resection			x	x
21	Splenectomy			x	x
22	Stomach procedures			x	x
23	Thyroid procedures	x	x		
24	Ulcer surgery			x	x
25	Umbilical hernia procedures	x			x
26	Ventral hernia repair	x	x		

Table 5: Mortality in 23,715 matched pairs of a patient receiving surgery at a magnet hospital or a control hospital, where the pairs have been divided into five groups selected by CART.

	Subgroups					Pooled
	Group 1	Group 2	Group 3	Group 4	Group5	
CHF	no	no	no	yes	yes	
Procedures	proc1	proc2	proc2	proc3	proc4	
ER admission	both	no	yes	both	both	
Number of Pairs	10127	5636	2943	2086	2923	23715
Discordant Pairs	210	293	488	217	760	1968
Percent Discordant %	2.1	5.2	16.6	10.4	26.0	8.3
Odds Ratio	1.41	1.53	1.09	1.28	1.18	1.23
Mortality %, Magnet	0.9	2.5	10.1	4.9	16.5	4.7
Mortality %, Control	1.3	3.5	10.8	6.2	18.6	5.6
Sensitivity analysis: Upper bounds on P-values for various Γ						
Γ	Subgroups					Truncated
	Group 1	Group 2	Group 3	Group 4	Group 5	Product
1.00	0.008	0.000	0.195	0.039	0.013	0.000
1.05	0.019	0.001	0.374	0.080	0.062	0.000
1.10	0.042	0.003	0.576	0.143	0.184	0.012
1.15	0.079	0.010	0.753	0.230	0.386	0.032
1.17	0.099	0.015	0.809	0.270	0.479	0.044
1.20	0.135	0.025	0.875	0.335	0.616	0.163

Table 6: Use of the intensive care unit (ICU) in 23,715 matched pairs of a patient receiving surgery at a magnet hospital or a control hospital, where the pairs have been divided into five groups indicated in Figure 4.

	Subgroups					Pooled
	Group 1	Group 2	Group 3	Group 4	Group5	
CHF	no	no	no	yes	yes	
Procedures	proc1	proc2	proc2	proc3	proc4	
ER admission	both	no	yes	both	both	
Number of Pairs	10127	5636	2943	2086	2923	23715
Discordant Pairs	2675	2361	1282	859	970	8147
Percent Discordant %	26.4	41.9	43.6	41.2	33.2	34.4
Odds ratio	1.63	2.05	1.67	1.70	1.88	1.78
ICU %, Magnet	15.3	28.9	53.8	41.0	69.8	32.3
ICU %, Control	21.7	43.3	64.6	51.7	80.0	42.0
Sensitivity analysis: Upper bounds on P-values for various Γ						
Γ	Subgroups					Truncated
	Group 1	Group 2	Group 3	Group 4	Group 5	Product
1	0.000	0.000	0.000	0.000	0.000	0.000
1.5	0.017	0.000	0.037	0.040	0.000	0.000
1.6	0.312	0.000	0.254	0.203	0.009	0.000
1.7	0.849	0.000	0.651	0.511	0.074	0.000
1.8	0.993	0.002	0.916	0.798	0.276	0.049
1.9	1.000	0.047	0.989	0.945	0.582	0.235

Table 7: Mortality and ICU use in 5,636 pairs in Group 2. The table counts pairs of patients, not individual patients.

Control Hospital	Magnet Hospital			
	Dead	Alive, ICU	Alive, no ICU	Total
Dead	23	72	105	200
Alive, ICU	60	744	1493	2297
Alive, no ICU	56	726	2357	3139
Total	139	1542	3955	5636

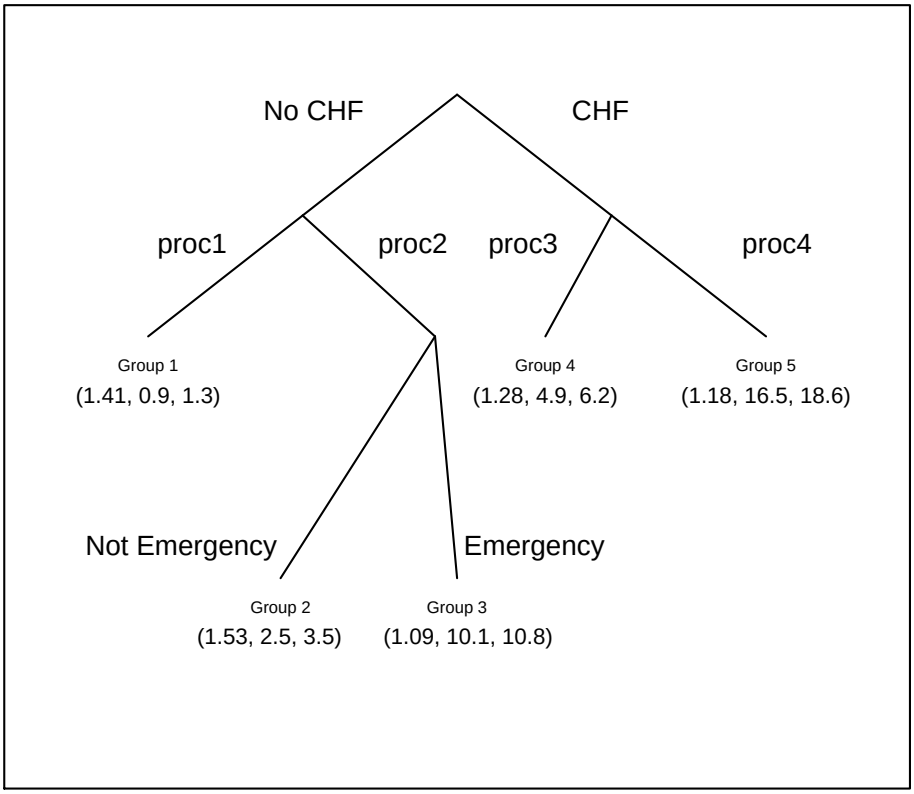

Figure 4: Mortality in 23,715 matched pairs of two Medicare patients, one receiving surgery at a magnet hospital identified for superior nursing, the other undergoing the same surgical procedure at a conventional control hospital. The three values (A,B,C) at the nodes of the tree are: A = McNemar odds ratio for mortality, control/magnet, B = 30-day mortality rate (%) at the magnet hospitals, C = 30-day mortality rate (%) at the control hospitals.

CHAPTER 4 : A New, Powerful Approach to the Study of Effect Modification in Observational Studies

4.1. Does Physical Activity Prolong Life? Equally for Everyone?

4.1.1. A Matched Comparison of Physical Inactivity and Survival

Davis et al. (1994) used the NHANES I Epidemiologic follow-up study (NHEFS) to ask: Is greater physical activity reported at the time of the NHANES I survey associated with a longer subsequent life? We examine the same data in a similar way, but with new methodology, specifically the subgroup maximum method or submax-method.

A representative national sample was collected in the first NHANES I survey in 1971-1975, and these sampled individuals were followed up for survival until 1992. Data on all variables other than death were collected at baseline (NHANES I). Physical activity was measured in two variables: self-reported nonrecreational activity ("In your usual days, aside from recreation are you physically very active, moderately active or quite inactive?") and self-reported recreational activity ("Do you get much, moderate or little or no exercise in the things you do for recreation?"). We formed a treated group of 470 adults who were quite inactive, both at work and at leisure, and we matched them to a control group of 470 adults who were quite active (very active in physical activity outside of recreation and much or moderate recreational activity). We compare quite inactive to quite active rather than moderately inactive to moderately active individuals because making the treated and control groups sharply differ in dose increases the insensitivity of the study to unobserved biases when there is a treatment effect and no bias (i.e., it increases the design sensitivity, Rosenbaum (2004)). Following Davis et al. (1994), we excluded people who were quite ill at the time of the NHANES I survey. Both of our groups included people aged between 45 and 74 at baseline in the NHANES I study and excluded people who, prior to the NHANES I evaluation, had had heart failure, a heart attack, stroke, diabetes, polio or paralysis, a malignant tumor, or a fracture of the hip or spine. Table 8 shows the covariates used in matching. Pairs were exactly matched on sex, smoking status (current smoker) and income (cut at two times the Federal poverty level). Other matched variables were age, race (white or other), years of education, employment (employed or not employed outside the home during the previous three months), marital status, alcohol consumption and dietary quality (number of five nutrients – protein, calcium, iron, Vitamin A and Vitamin C – that were consumed at more than two thirds of the recommended dietary allowance). After matching, the groups are fairly similar, whereas before matching, the inactive group was older, more often female, more often nonwhite, more often poor, more often not working in the prior 3 months, more often not married, and less often had an adequate diet.

The top panel of Figure 5 shows the Kaplan-Meier survival curves for the matched active and inactive groups. We ask two interconnected questions: (i) What magnitude of unmeasured bias from nonrandom treatment assignment would need to be present to explain Figure 5 as something other than an effect caused by inactivity? (ii) Is there greater insensitivity to unmeasured bias in some subgroups because the ostensible effect is larger in those subgroups, or is there similar evidence of effect in all subgroups? We will study sex, smoking and the

two categories of income as potential effect modifiers. These three binary covariates are exactly matched.

4.1.2. A New Approach to Effect Modification in Observational Studies

If some subgroups experience larger or more stable effects, then the ostensible effect of a treatment may be less sensitive to bias from nonrandomized treatment assignment in these subgroups; see Hsu et al. (2013). Conversely, if a treatment appears to be highly effective in all subgroups, then it is safer to generalize to other populations that may have different proportions of people in the various subgroups.

One approach to studying effect modification in observational studies constructs a few promising subgroups from several measured covariates using an algorithm such as Breiman et al. (1984)'s CART technique, as discussed by Hsu et al. (2013, 2015), and as described in §4.3.7. A limitation of this approach is that it is hard to study the power and operating characteristics of such a technique except by simulation, because the CART step does not lend itself to such an evaluation. In the current paper we propose a different approach — the submax method — for which a theoretical evaluation is possible. The submax method has a formula for power and design sensitivity, and additionally permits statements about Bahadur efficiency. In particular, the new method achieves the largest — i.e., best — of the design sensitivities for the subgroups, and the highest Bahadur efficiency of the subgroups; moreover, both the power formula and a simulation confirm that the asymptotic results are a reasonable guide to performance in samples of practical size. The simulation in §4.3.7 also compares the submax and CART methods. An additional limitation of the CART method is that it is defined for matched pairs. In contrast, the submax-method works for matched pairs, for matched sets with multiple controls, variable numbers of controls and with the full matching method described by Rosenbaum (1991) and Hansen and Klopfer (2006).

The submax-method considers a single combined analysis together with several ways to split the population into subgroups. It does not form the interaction of subgroups, which would quickly become thinly populated with small sample sizes; rather, it considers one split, reassembles the population, then considers another split. If the splits were defined by L binary covariates, then there would be 2^L interaction subgroups, but the submax-method would do only 1 overall test plus $2L$ subgroup tests, making a total of $2L+1$ highly correlated tests, not 2^L independent tests. If the binary covariates each split every subpopulation in half, then each interaction subgroup would contain a fraction 2^{-L} of the population — i.e., not much — but each of our $2L$ subgroup tests would use half the population — i.e., a much larger fraction. The submax-method uses the joint distribution of the $2L + 1$ test statistics, with the consequence that the correction for multiple testing is quite small due to the high correlation among the test statistics. Specifically, the two halves of one binary split are independent because they refer to different people, but each of those test statistics is highly correlated with test statistics for other splits, because all the splits use the same people. In the example, we split the population by gender (male or female), by current cigarette smoking (yes or no), and by two income groups, so we do $2K + 1 = 2 \times 3 + 1 = 7$ correlated tests. Although the test statistics for men and women are independent, the

statistics for men and smokers are highly correlated because there are many male smokers.

4.2. Notation and Review of Observational Studies

4.2.1. Treatment Effects in Randomized Experiments

There are G groups, $g = 1, \ldots, G$, of matched sets, $i = 1, \ldots, I_g$, with n_{gi} individuals in set i, $j = 1, \ldots, n_{gi}$, one treated individual with $Z_{gij} = 1$ and $n_{gi} - 1$ controls with $Z_{gij} = 0$, so that $1 = \sum_{j=1}^{n_{gi}} Z_{gij}$ for each g, i. Matched sets were formed by matching for an observed covariate x_{gij}, but may fail to control an unobserved covariate u_{gij}, so that $x_{gij} = x_{gik}$ for each g, i, j, k, but possibly $u_{gij} \neq u_{gik}$. In §4.1.1, the matched sets are pairs, $n_{gi} = 2$, and there are $G = 2^3 = 8$ groups of pairs defined by combinations of $L = 3$ binary covariates, sex, smoking and income group, with $470 = \sum_{g=1}^{8} I_g$ pairs in total.

Individual gij exhibits response r_{Tgij} if treated or response r_{Cgij} if given the control, so this individual exhibits response $R_{gij} = Z_{gij}\, r_{Tgij} + (1 - Z_{gij})\, r_{Cgij}$, and the effect of the treatment, $r_{Tgij} - r_{Cgij}$, is not observed for anyone; see Neyman (1923, 1990) and Rubin (1974). Fisher's (1935) null hypothesis of no treatment effect asserts that $H_0 : r_{Tgij} = r_{Cgij}$ for all i, j. Write $\mathcal{F} = \{(r_{Tgij}, r_{Cgij}, x_{gij}, u_{gij}),\, g = 1, \ldots, G,\, i = 1, \ldots, I_g,\, j = 1, \ldots, n_{gi}\}$. Write $|\mathcal{S}|$ for the number of elements in a finite set $\mathcal{S}$.

Write $\mathcal{Z}$ for the set containing the $|\mathcal{Z}| = \prod_{g=1}^{G} \prod_{i=1}^{I_g} n_{gi}$ possible values $\mathbf{z}$ of the treatment assignment $\mathbf{Z} = \left(Z_{111}, Z_{112}, \ldots, Z_{G,I_G,n_{G,I_G}}\right)^T$, so $\mathbf{z} \in \mathcal{Z}$ if $z_{gij} = 0$ or $z_{gij} = 1$ and $1 = \sum_{j=1}^{n_{gi}} z_{gij}$ for each gi. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. In an experiment, randomization picks a $\mathbf{Z}$ at random from $\mathcal{Z}$, so that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}$. In a randomized experiment, randomization creates the exact null randomization distribution of familiar test statistics, such as Wilcoxon's signed rank statistic or the mean pair difference or Maritz (1979)'s version of Huber M-statistic. In the analysis of the paired censored survival data in §4.1.1, the test statistic is the Prentice-Wilcoxon test proposed by O'Brien and Fleming (1987). These test statistics and many others are of the form $T = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \sum_{j=1}^{n_{gi}} Z_{gij}\, q_{gij}$ for suitable scores q_{gij} that are a function of the R_{gij}, n_{gi} and possibly the x_{gij}, so that, under H_0 in a randomized experiment, the conditional distribution $\Pr(T \mid \mathcal{F}, \mathcal{Z})$ of the test statistic T is the distribution of the sum of fixed scores q_{gij} with $Z_{gij} = 1$ selected at random. In a conventional way, randomization tests are inverted to obtain confidence intervals and point estimates for magnitudes of treatment effects; see, for instance, Lehmann (1975), Maritz (1979) and Rosenbaum (2007).

In large sample approximations, the number of groups, G, will remain fixed, and the number of matched sets I_g in each group will increase without bound.

4.2.2. Sensitivity to Unmeasured Biases in Observational Studies

In an observational study, conventional tests of H_0 appropriate in the randomized experiments in §4.2.1 can falsely reject a true null hypothesis of no effect because treatments are not assigned at random, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) \neq |\mathcal{Z}|^{-1}$. A simple model for sensitivity analysis in observational studies assumes that, in the population prior to matching for x, treatment

assignments are independent and two individuals, gij and $g'i'j'$, with the same observed covariates, $x_{gij} = x_{g'i'j'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\Pr\left(Z_{gij} = 1 \,|\, \mathcal{F}\right) \Pr\left(Z_{g'i'j'} = 0 \,\middle|\, \mathcal{F}\right)}{\Pr\left(Z_{g'i'j'} = 1 \,\middle|\, \mathcal{F}\right) \Pr\left(Z_{gij} = 0 \,|\, \mathcal{F}\right)} \leq \Gamma \text{ whenever } x_{gij} = x_{g'i'j'}; \tag{4.2.1}$$

then the distribution of $\mathbf{Z}$ is returned to $\mathcal{Z}$ by conditioning on $\mathbf{Z} \in \mathcal{Z}$.

Under the model (4.2.1), one obtains conventional randomization inferences for $\Gamma = 1$, but these are replaced by an interval of P-values or an interval of point estimates or an interval of endpoints for a confidence interval for $\Gamma > 1$. The intervals become longer as Γ increases, the interval of P-values tending to $[0, 1]$ as $\Gamma \to \infty$, reflecting the familiar fact that association, no matter how strong, does not logically entail causation. At some point, the interval is sufficiently long to be uninformative, for instance including P-values that would both reject and accept the null hypothesis of no effect. The question answered by a sensitivity analysis is: How much bias in treatment assignment, measured by Γ, would need to be present before the study becomes uninformative? For instance, how large would Γ have to be to produce a P-value above α, conventionally $\alpha = 0.05$?

An approximation to the upper bound on the P-value is obtained in the following way; see Gastwirth et al. (2000) for detailed discussion and see Rosenbaum (2007) for its application to Huber-Maritz M-tests. Assume H_0 is true for the purpose of testing it, so that $R_{gij} = r_{Cgij}$ and q_{gij} are fixed by conditioning on $\mathcal{F}$. Write $T_g = \sum_{i=1}^{I_g} \sum_{j=1}^{n_{gi}} Z_{gij}\, q_{gij}$, so that $T = \sum_{g=1}^{G} T_g$. Subject to (4.2.1) for a given $\Gamma \geq 1$, find the maximum expectation, $\mu_{\Gamma g}$, of T_g. Also, among all treatment assignment probabilities that satisfy (4.2.1) and that achieve the maximum expectation $\mu_{\Gamma g}$, find the maximum variance, $\nu_{\Gamma g}$, of T_g. If $T \geq \sum_{g=1}^{G} \mu_{\Gamma g}$, report as the upper bound on the P-value for T,

$$1 - \Phi\left\{\left(\sum_{g=1}^{G} T_g - \mu_{\Gamma g}\right) \Big/ \sqrt{\sum_{g=1}^{G} \nu_{\Gamma g}}\right\}, \tag{4.2.2}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution. The bound is derived as $\min(I_g) \to \infty$ with some mild conditions to ensure that no one q_{gij} dominates the rest, and that the fixed scores q_{gij} do not become degenerate as $\min(I_g)$ increases. For $\Gamma = 1$, this yields a Normal approximation to a randomization P-value using T as the test statistic. If treatment assignments were governed by the probabilities satisfying (4.2.1) that yield $\mu_{\Gamma g}$ and $\nu_{\Gamma g}$, then, under H_0 and mild conditions on the q_{gij}, the joint distribution of

$$\left\{\left(T_1 - \mu_{\Gamma 1}\right) / \sqrt{\nu_{\Gamma 1}}, \ldots, \left(T_G - \mu_{\Gamma G}\right) / \sqrt{\nu_{\Gamma G}}\right\}^T$$

would converge to a G-dimensional Normal distribution with expectation vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$ as $\min(I_g) \to \infty$. Simpler methods of proof and formulas apply in simple cases, such as matched pairs; for instance, contrast §3 and §4 of Rosenbaum (2007). These simpler methods of proof bound the distribution of T exactly, then approximate the bounding distribution, whereas the general method is merely a large sample approximation

to the upper bound on the P-value when $T \geq \sum_{g=1}^{G} \mu_{\Gamma g}$. Write $\boldsymbol{\mu}_\Gamma = (\mu_{\Gamma 1}, \ldots \mu_{\Gamma G})^T$ and $\mathbf{V}_\Gamma$ for the $G \times G$ diagonal matrix with gth diagonal element $\nu_{\Gamma g}$.

For various methods of sensitivity analysis in observational studies, see Egleston et al. (2009), Gilbert et al. (2003), Hosman et al. (2010), and Liu et al. (2013).

4.2.3. Design Sensitivity and Bahadur Efficiency

Suppose that there is a treatment effect and there is no bias from the unobserved covariate u_{gij}, and call this the favorable situation. In an observational study, if an investigator were in the favorable situation, then she would not know it, and the best she could hope to say is that the results are insensitive to small and moderate biases Γ. The power of a sensitivity analysis is the probability that she will be able to say this. More precisely, in the favorable situation, the power of a level α sensitivity analysis at sensitivity parameter Γ is the probability that (4.2.2) will be less than or equal to α when computed at the given Γ.

As the sample size increases, there is a value, $\widetilde{\Gamma}$, called the design sensitivity, such that the power tends to 1 if $\Gamma < \widetilde{\Gamma}$ and the power tends to zero if $\Gamma > \widetilde{\Gamma}$, so $\widetilde{\Gamma}$ is the limiting sensitivity to unmeasured bias for a given favorable situation and test statistic; see Rosenbaum (2004); Rosenbaum (2010), Zubizarreta et al. (2013) and Stuart and Hanna (2013). In a particular favorable situation, for a specific Γ, the rate at which (4.2.2) declines to zero with increasing sample size yields the Bahadur efficiency of the sensitivity analysis, and the efficiency drops to zero at $\Gamma = \widetilde{\Gamma}$; see Rosenbaum (2015).

4.3. Joint Bounds for Two or More Comparisons

4.3.1. Subgroup Comparisons

We are interested in K specified comparisons, $k = 1, \ldots, K$, among the G groups of matched sets. By one comparison we mean a fixed nonzero vector $\mathbf{c}_k = (c_{1k}, \ldots, c_{Gk})^T$ of dimension G with $c_{gk} \geq 0$ for $g = 1, \ldots, G$, and we evaluate a comparison using the statistic $S_k = \sum_{g=1}^{G} c_{gk} T_g$. For instance, the comparison $\mathbf{c}_1 = (1, \ldots, 1)^T$ yields the overall test in §4.2.2. By replacing the scores q_{gij} in §4.2.2 by scores $q^*_{gij} = c_{gk}\, q_{gij}$, the bound for S_k is obtained in parallel with (4.2.2). If groups $1, \ldots, G/2$ are matched sets of men and groups $G/2+1, \ldots, G$ are matched sets of women, then the comparison $\mathbf{c}_2 = (1, \ldots, 1, 0, \ldots, 0)^T$ confines attention to men, while the comparison $\mathbf{c}_3 = (0, \ldots, 0, 1, \ldots, 1)^T$ confines attention to women. Perhaps an additional comparison $\mathbf{c}_4 = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^T$ would confine attention to people over the age of 65, and so on.

If the treatment effect for women were larger than the effect for men, the comparison, $\mathbf{c}_3$, restricted to women might be insensitive to larger unmeasured biases than the overall comparison, $\mathbf{c}_1$. Hsu et al. (2013) present an example in which a treatment to prevent malaria is far more effective for children than for adults, so that only very large biases in treatment assignment could explain away the ostensible benefits for children.

4.3.2. Joint Evaluation of Subgroup Comparisons

Let $\mathbf{C}$ be the $K \times G$ matrix whose K rows are the $\mathbf{c}_k^T = (c_{1k}, \ldots, c_{Gk})$, $k = 1, \ldots, K$. Define $\boldsymbol{\theta}_\Gamma = \mathbf{C}\boldsymbol{\mu}_\Gamma$ and $\boldsymbol{\Sigma}_\Gamma = \mathbf{C}\mathbf{V}_\Gamma\mathbf{C}^T$, noting that $\boldsymbol{\Sigma}_\Gamma$ is not typically diagonal. Write $\theta_{\Gamma k}$ for the kth coordinate of $\boldsymbol{\theta}_\Gamma$ and $\sigma^2_{\Gamma k}$ for the kth diagonal element of $\boldsymbol{\Sigma}_\Gamma$. Define $D_{\Gamma k} = (S_k - \theta_{\Gamma k})/\sigma_{\Gamma k}$ and $\mathbf{D}_\Gamma = (D_{\Gamma 1}, \ldots, D_{\Gamma K})^T$. Finally, write $\boldsymbol{\rho}_\Gamma$ for the $K \times K$ correlation matrix formed by dividing the element of $\boldsymbol{\Sigma}_\Gamma$ in row k and column k' by $\sigma_{\Gamma k}\,\sigma_{\Gamma k'}$. Subject to (4.2.1) under H_0, at the treatment assignment probabilities that yield the $\mu_{\Gamma g}$ and $\nu_{\Gamma g}$, the distribution of $\mathbf{D}_\Gamma$ is converging to a Normal distribution, $N_K(\mathbf{0}, \boldsymbol{\rho}_\Gamma)$, with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{\rho}_\Gamma$ as $\min(I_g) \to \infty$. Using this null distribution, the null hypothesis H_0 is tested using

$$D_{\Gamma\,\max} = \max_{1\le k\le K} D_{\Gamma k} = \max_{1\le k\le K} \frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}}.$$

The α critical value $\kappa_{\Gamma,\alpha}$ for $D_{\Gamma\,\max}$ solves

$$1 - \alpha = \Pr(D_{\Gamma\,\max} < \kappa_{\Gamma,\alpha}) = \Pr\left(\frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}} < \kappa_{\Gamma,\alpha},\ k = 1, \ldots, K\right) \tag{4.3.1}$$

under H_0. The multivariate Normal approximation to $\kappa_{\Gamma,\alpha}$ is obtained using the `qmvnorm` function in the `mvtnorm` package in R, as applied to the $N_K(\mathbf{0}, \boldsymbol{\rho}_\Gamma)$ distribution; see Genz and Bretz (2009). Notice that this approximation to $\kappa_{\Gamma,\alpha}$ depends upon Γ only through $\boldsymbol{\rho}_\Gamma$, which in turn depends upon Γ only through $\nu_{\Gamma g}$. The resulting approximate α critical value $\kappa_{\Gamma,\alpha}$ for $D_{\Gamma\,\max}$ is larger than $\Phi^{-1}(1-\alpha)$ because the largest of K statistics $D_{\Gamma k}$ has been selected, and it reflects the correlations $\boldsymbol{\rho}_\Gamma$ among the coordinates of $\mathbf{D}_\Gamma$.

4.3.3. Behavior of the critical constant $\kappa_{\Gamma,\alpha}$ in a simple case

Consider a simple, balanced case under the null hypothesis H_0, in which every matched set is a matched pair, $n_{gi} = 2$ for all g, i, and outcomes are continuously distributed and hence untied with probability one. Additionally, there are L matched binary covariates, such as gender, to be examined as potential effect modifiers making $G = 2^L$ groups of pairs, with the same number of matched pairs in each group, $I_1 = \cdots = I_G = \overline{I}$, say. Suppose that, in each group, T_g is Wilcoxon's signed rank statistic computed from the $\overline{I}$ pairs in that group. In this case, $\mu_{\Gamma g} = \{\Gamma/(1+\Gamma)\}\,\overline{I}\,(\overline{I}+1)/2$ and $\nu_{\Gamma g} = \left\{\Gamma/(1+\Gamma)^2\right\}\overline{I}\,(\overline{I}+1)\,(2\overline{I}+1)/6$; see (Rosenbaum, 2002b), §4.3.3. In this simple case, by symmetry, the correlation matrix $\boldsymbol{\rho}_\Gamma$ does not depend upon Γ. There are $K = 2L+1$ comparisons, namely $\mathbf{c}_1 = (1, \ldots, 1)^T$ in §4.3.1 using all of the pairs, yielding T as in §4.2.2, plus two comparisons for each binary covariate for half the pairs at the high and low levels of that covariate, for instance, $\mathbf{c}_2$, $\mathbf{c}_3$ and $\mathbf{c}_4$ in §4.3.1, making a total of $K = 2L+1$ tests. Because of the symmetry of this situation, the correlation/covariance matrix $\boldsymbol{\rho}_\Gamma$ of $D_{\Gamma k}$ has the simple form in Table 9; that is, $D_{\Gamma 1}$ has correlation $0.707 = 1/\sqrt{2}$ with $D_{\Gamma k}$ for $k \ge 2$, the two consecutive comparisons for the two categories of the same binary variable are uncorrelated, and all other comparisons have correlation 0.5.

In this simple, balanced case, Table 10 shows the critical constant $\kappa_{\Gamma,\alpha}$ for $\alpha = 0.05$ and $L =$

0, 1, ..., 15 potential effect modifiers, and $K = 2L+1 = 1, 3, \ldots, 31$ tests. For comparison in Table 10, $\kappa_{\Gamma,\alpha}$ is compared to $\Phi^{-1}(1-\alpha/K)$, the critical constant obtained from the Bonferroni inequality. For instance, the Bonferroni critical constant $\Phi^{-1}(1-\alpha/K)$ for $K = 15$ tests and $L = 7$ is 2.71, which is larger than the submax critical constant of 2.70 for $K = 25$ tests and $L = 12$.

4.3.4. Application in the NHANES Example

Table 11 performs the test in §4.3.2 for the NHANES data in §4.1.1 using the Prentice-Wilcoxon statistic T of O'Brien and Fleming (1987). The row of Table 11 for $\Gamma = 1$ consists of Normal approximations to randomization tests, while the rows with $\Gamma > 1$ examine sensitivity to bias from nonrandom treatment assignment. For $\Gamma = 1$, the test statistic $D_{\Gamma\,\max} = 6.09 \geq \kappa_{\Gamma,\alpha} = 2.31$, so Fisher's hypothesis of no treatment effect would be rejected at level α if the data had come from a randomized experiment with $\Gamma = 1$. For $\Gamma = 1$, the maximum statistic is based on all 470 pairs, $D_{\Gamma\,\max} = D_{\Gamma 1}$; however, $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha} = 2.31$ for every subgroup, $k = 1, \ldots, K = 7$. At $\Gamma = 1.4$, the deviates $D_{\Gamma 2}$ and $D_{\Gamma 6}$ for females ($k = 2$) and the nonpoor ($k = 6$) no longer exceed $\kappa_{\Gamma,\alpha} = 2.31$, and the precise meaning of this is examined in more detail in §4.4. At $\Gamma = 1.64$, Fisher's hypothesis of no treatment effect is still rejected because the deviate $D_{\Gamma 3}$ for males exceeds $\kappa_{\Gamma,\alpha} = 2.31$. Although there are 275 pairs of women and 195 pairs of men, the strongest evidence, the least sensitive evidence, of an effect of inactivity on survival is for men. The bottom two panels of Figure 5 show the separate survival curves for men and women.

Table 11 is compactly and conveniently indexed by one parameter Γ. It is sometimes helpful to give a two-parameter interpretation of this one parameter. In particular, the longer life of active men in Table 11 is insensitive to an unmeasured bias of $\Gamma = 1.64$. In a matched pair, $\Gamma = 1.64$ corresponds with an unobserved covariate that doubles the odds of a longer life and increases the chance of inactivity by a factor of more than 6-fold; see the amplification of Γ into two equivalent parameters Δ and Λ in Rosenbaum and Silber (2009a), where $1.64 = \Gamma = (\Delta\Lambda + 1)/(\Delta\Lambda)$ for $\Delta = 2$ and $\Lambda = 6.33$.

In §4.3.8, an alternative analysis of the NHANES data is presented using Breiman et al. (1984)'s CART regression, as proposed by Hsu et al. (2013, 2015). The CART technique is described in §4.3.7 where a simulation compares it to the submax method.

4.3.5. Design Sensitivity and Bahadur Efficiency

As in Rosenbaum (2012), it is easy to see that under an alternative hypothesis given by a favorable situation — a treatment effect with no unmeasured bias — the design sensitivity of $D_{\Gamma\,\max}$, say $\widetilde{\Gamma}_{\max}$, is equal to the maximum design sensitivity $\widetilde{\Gamma}_k$ of the K component tests, $\widetilde{\Gamma}_{\max} = \max\left(\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_K\right)$. Briefly, by the definition of design sensitivity, if $\Gamma < \widetilde{\Gamma}_k$, then the probability that $D_{\Gamma k} \geq \kappa$ tends to 1 for every κ as $\min(I_g) \rightarrow \infty$, so the probability that $D_{\Gamma\,\max} \geq \kappa_{\Gamma,\alpha}$ tends to 1 because $D_{\Gamma\,\max} \geq D_{\Gamma k}$. Although there is a price to be paid for multiple testing, that price does not affect the design sensitivity.

Define $\beta_1 = 1$. Berk and Jones (1979) show that, if $D_{\Gamma k}$ has Bahadur efficiency β_k relative to $D_{\Gamma 1}$ for $k = 2, \ldots, K$ under some alternative hypothesis, then $D_{\Gamma\,\max}$ has Bahadur

efficiency $\beta_{\max} = \max_{1 \le k \le K} \beta_k$. Berk and Jones call this "relative optimality" meaning $D_{\Gamma\,\max}$ is optimal among the fixed set $D_{\Gamma 1}, \ldots, D_{\Gamma K}$. In other words, the correction for multiplicity, $\kappa_{\Gamma,\alpha} > \Phi^{-1}(1-\alpha)$, does reduce finite sample power, but in a limited way, so that the Bahadur efficiency is ultimately unaffected.

4.3.6. Power Calculations and Design Sensitivity in a Simple Case

Under an alternative hypothesis, if the T_g are independent and asymptotically Normal with expectation μ_g^* and variance ν_g^*, then straightforward manipulations involving the multivariate Normal distribution yield an asymptotic approximation to the power of tests based on $D_{\Gamma\,\max}$ or $D_{\Gamma k}$ for fixed k.

Specifically, write $\theta_k^* = \sum_{g=1}^G c_{gk}\,\mu_g^*$ and σ_k^* for the square root of the kth diagonal element of the $K \times K$ covariance matrix $\mathbf{C}\text{diag}\left(\nu_1^*, \ldots, \nu_K^*\right)\mathbf{C}^T$, so θ_k^* is the expectation and σ_k^* is the standard deviation of S_k under the alternative; moreover, write $\boldsymbol{\rho}^*$ for the $K \times K$ correlation matrix computed from this covariance matrix. The approximate power is the following probability computed under the alternative hypothesis,

$$1 - \Pr\left(D_{\Gamma\,\max} < \kappa_{\Gamma,\alpha}\right) = 1 - \Pr\left(\frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}} < \kappa_{\Gamma,\alpha},\, k = 1, \ldots, K\right)$$

$$= 1 - \Pr\left(\frac{S_k - \theta_k^*}{\sigma_k^*} < \frac{\theta_{\Gamma k} - \theta_k^* + \kappa_{\Gamma,\alpha}\,\sigma_{\Gamma k}}{\sigma_k^*},\, k = 1, \ldots, K\right). \tag{4.3.2}$$

The Normal approximation to the joint distribution of the T_g under the alternative means that the last term in (4.3.2) is approximately a particular quadrant probability for the $N_K\left(\mathbf{0}, \boldsymbol{\rho}^*\right)$ distribution, and this may be calculated using the `pmvnorm` function in the `mvtnorm` package in `R`. Under the same assumptions, the power of a test based on one fixed $D_{\Gamma k}$ is approximately

$$1 - \Pr\left\{\frac{S_k - \theta_k^*}{\sigma_k^*} < \frac{\theta_{\Gamma k} - \theta_k^* + \Phi^{-1}(1-\alpha)\;\sigma_{\Gamma k}}{\sigma_k^*}\right\}, \tag{4.3.3}$$

and this may be calculated using the standard Normal distribution.

Moreover, the design sensitivity $\widetilde{\Gamma}_k$ for $S_k = \sum_{g=1}^G c_{gk} T_g$ is the limit of values of Γ that solve $1 = \left(\sum_{g=1}^G c_{gk}\,\mu_g^*\right) / \left(\sum_{g=1}^G c_{gk}\,\mu_{\Gamma g}\right)$. That is, using S_k, as $I \to \infty$, the power tends to 1 for $\Gamma < \widetilde{\Gamma}_k$ and it tends to 0 for $\Gamma > \widetilde{\Gamma}_k$. This formula emphasizes the importance of effect modification. For instance, with two groups, $G = 2$, say $g = 0$ and $g = 1$, if $\mu_0^* > \mu_1^*$, then the design sensitivity is largest with $c_{0k} = 1$ and $c_{1k} = 0$, so as $I \to \infty$, there are values of Γ such that the power of the overall test is tending to 0 while the power of a test focused on the first subgroup is tending to 1. This will be quite visible in both theoretical and simulated power calculations.

An oracle would use the one $D_{\Gamma k}$ with the highest power. Lacking such an oracle, it is interesting to compare $D_{\Gamma\,\max}$ to: (i) the oracle, (ii) the one test, $D_{\Gamma 1}$, that uses all of the

matched sets, as in §4.2.2.

To illustrate, consider the simple, balanced case in §4.3.3, and suppose that there are L binary covariates as potential effect modifiers. We would like to compute power under a favorable alternative, meaning that, unknown to the investigator, the treatment has an effect and there is no unmeasured bias from u_{gij}. Because the investigator cannot know that the data came from the favorable situation, a sensitivity analysis is performed. A simple favorable situation has $I_g = \overline{I}$ independent treated-minus-control pair differences in every group g, where the pair differences are Normal with various expectations and variance 1. Then Wilcoxon's signed rank statistic in group g, namely T_g, is asymptotically Normal under the alternative hypothesis as $\overline{I} \rightarrow \infty$, and simple formulas in Lehmann (1975), §4.2 give the expectation and variance, μ_g^* and ν_g^*, of T_g, under this alternative. There are $G\overline{I} = 2^L \cdot \overline{I}$ pairs in total. Note that the $K = 2L + 1$ statistics, S_k, are each computed from at least $2^{L-1} \cdot \overline{I}$ pairs, not from $\overline{I}$ pairs, and they are each sums of at least 2^{L-1} signed rank statistics T_g.

Table 12 displays theoretical power for a level $\alpha = 0.05$ test of no effect in several favorable situations, that is, situations with a treatment effect and no bias. In Table 12, "one covariate" refers to $L = 1$ binary covariate, making $G = 2^L = 2$ groups, so that $D_{\Gamma\,\max}$ is the maximum of three statistics, namely the deviates for the signed rank statistics in groups 1 and 2 and for the sum of these two statistics. In Table 12, "five covariates" refers to $L = 5$ binary covariates, making $G = 2^L = 32$ groups, so that $D_{\Gamma\,\max}$ is the maximum of $11 = 2 \times 5 + 1$ statistics, namely the deviates for 10 totals of 16 signed rank statistics at the high and low levels of each covariate, and also for the sum of all 32 signed rank statistics.

The sample size in Table 12 is constant in each group, $I_g = \overline{I}$, with total sample size $2016 = G\overline{I} = 2^L \cdot \overline{I}$, so this is $\overline{I} = 1008$ for $L = 1$ covariate and $\overline{I} = 63$ for $L = 5$ covariates. In both cases, $L = 1$ and $L = 5$, only the first covariate is a potential effect modifier, in the sense that the expected pair difference only changes with the level of the first covariate, being ζ_0 for the 0 level and ζ_1 for the 1 level. When $\zeta_0 \neq \zeta_1$, there is effect modification. With $L = 5$, four of the five covariates are simply a distraction that require $D_{\Gamma\,\max}$ to make a larger correction for multiple testing. The first situation in Table 12 has no treatment effect, $\zeta_0 = \zeta_1 = 0$, so the reported values are the actual size of a level $\alpha = 0.05$ test. The second situation in Table 12 has a constant treatment effect, $\zeta_0 = \zeta_1 = 0.5$, so it is a mistake to look for effect modification because there is none. The third situation in Table 12 has slight effect modification, $\zeta_0 = 0.6 > 0.4 = \zeta_1$, although the average treatment effect is $0.5 = (\zeta_0 + \zeta_1)/2$ as in the second situation. The fourth situation in Table 12 has substantial effect modification, $\zeta_0 = 0.5 > 0 = \zeta_1$, so the average treatment effect is $0.25 = (\zeta_0 + \zeta_1)/2$. The design sensitivity $\widetilde{\Gamma}_g$ for Wilcoxon's statistic T_g in group g is 3.17 if $\zeta_g = 1/2$ and it is 1 if $\zeta_g = 0$; see Rosenbaum (2010, p. 272) for details of this calculation. For instance, in Table 12 with $\zeta_0 = \zeta_1 = 0.5$, the power of the test is below the level $\alpha = 0.05$ when $\Gamma > \widetilde{\Gamma}_g = 3.17$.

Table 12 compares the power of $D_{\Gamma\,\max}$ to a single combined test $D_{\Gamma 1}$ that uses all pairs and an oracle that performs a single test using all the pairs that have the largest value of ζ_g. Obviously, the oracle is not a statistical procedure because it requires the statistician

to know what she does not know, namely which groups have the largest ζ_g. From theory, in the nonnull situations 2, 3 and 4, we know that $D_{\Gamma\max}$ has the same design sensitivity as the oracle, whereas the $D_{\Gamma 1}$ has lower design sensitivity than the oracle unless there is no effect modification, $\zeta_0 = \zeta_1$, as in situation 2. In situation 2, all three procedures have design sensitivity $\widetilde{\Gamma} = 3.17$, with negligible power for $\Gamma = 3.2 > 3.17$. In situation 3, $\zeta_0 = 0.6$, and both $D_{\Gamma\max}$ and the oracle have design sensitivity $\widetilde{\Gamma} = 4.05$ by focusing on group 0 for covariate 1, and they have nonnegligible power at $\Gamma = 3.4 < 4.05$; however, $D_{\Gamma 1}$ has design sensitivity $\widetilde{\Gamma} = 3.13$ in situation 3, with negligible power at $\Gamma = 3.2$. In situation 4, $\zeta_1 = 0$, and both $D_{\Gamma\max}$ and the oracle have design sensitivity $\widetilde{\Gamma} = 3.17$ by focusing on group 0 for covariate 1; however, $D_{\Gamma 1}$ has design sensitivity $\widetilde{\Gamma} = 1.70$ in situation 3, with negligible power at $\Gamma = 2.8$.

In the first situation in Table 12, all tests have the correct size for $\Gamma = 1$, and because there is no actual bias in the favorable situation, they have size below 0.05 for $\Gamma > 1$. In the second situation in Table 12, $D_{\Gamma\max}$ pays a price in power in its search for effect modification that is not there. In situations 3 and 4, $D_{\Gamma\max}$ has much higher power than the $D_{\Gamma 1}$ statistic, but it is behind the oracle, reflecting the price paid to discover the true pattern of effect modification. For instance, at $\Gamma = 2.8$, with $L = 5$ binary covariates and slight effect modification, $\zeta_0 = 0.6 > 0.4 = \zeta_1$, the statistic $D_{\Gamma\max}$ has power .959, the oracle has power 0.996, and $D_{\Gamma 1}$ has power 0.521.

4.3.7. Simulated Power and a Comparison with CART Groups

Table 13 describes simulated power for the same situation as the theoretical power in Table 12. Unlike Table 12, the simulation includes the power for a competing method for matched pairs proposed by Hsu et al. (2015), in which groups are built from covariates using the CART procedure of Breiman et al. (1984). There is no known power formula for the CART method, so it cannot be included in Table 12. In this approach, the pairs are initially ungrouped, and so lack a g subscript. However, the pairs have been exactly matched for several covariates that may be effect modifiers. The absolute treated-minus-control pair difference in outcomes in pair i, namely $|Y_i| = |R_{i1} - R_{i2}|$, is regressed on these covariates using CART, and the leaves of the tree define the groups. The P-values with the groups so-defined are combined using the truncated product of P-values proposed by Zaykin et al. (2002). The truncated product is analogous to Fisher's product of P-values, except P-values above a prespecified truncation point, ς, enter the product as 1, so the two methods are the same for $\varsigma = 1$. In Table 12, $\varsigma = 1/10$. Unlike $D_{\Gamma\max}$, there is no guarantee that the CART procedure will equal the oracle in terms of design sensitivity. In other words, we expect $D_{\Gamma\max}$ to win in sufficiently large samples, tracking the oracle as $\min(I_g) \to \infty$; however, $D_{\Gamma\max}$ may not win in the finite samples in Table 13.

Table 13 provides a check on the theoretical power formulas that yielded Table 12, and in general the two tables are in agreement. The CART procedure has higher power than $D_{\Gamma\max}$ when there is no effect modification in situation 2, $\zeta_0 = \zeta_1 = 0.5$, because the CART procedure typically produces a single group in this situation. The CART procedure has lower power than $D_{\Gamma\max}$ when there is slight effect modification in situation 3, $\zeta_0 = .6 > .4 = \zeta_1$, perhaps because the CART procedure fails to locate the slight effect modification.

In situation 4, with $\zeta_0 = .5 > 0 = \zeta_1$, the move from $L = 1$ covariate to $L = 5$ covariates reduces the power of both $D_{\Gamma\,\max}$ and the CART procedure, but it does more harm to $D_{\Gamma\,\max}$. Presumably, $D_{\Gamma\,\max}$ pays a higher price for multiple testing with $L = 5$ than with $L = 1$ consistent with Table 10, while the CART procedure has more difficulty finding the right groups with $L = 5$ than with $L = 1$.

There is no uniform winner in Table 13. However, when compared to the CART method, we expect $D_{\Gamma\,\max}$ to gradually catch up, or to move ahead, or to stay ahead as $\min(I_g) \to \infty$ because it has the best design sensitivity; therefore, relative performance depends upon the sample size.

4.3.8. Use of CART in the Example

As an alternative to the analysis in §4.3.4, consider using the CART method in §4.3.7, implemented using the `rpart` package in `R`. In an `rpart` tree, the number of splits is controlled by a complexity parameter that defaults to the value 0.01. Using the default settings in `rpart`, the CART tree is a single group of all 470 pairs. At $\Gamma = 1.64$, the single group test has deviate $D_{\Gamma 1} = 2.29$ and one-sided P-value bound of $1 - \Phi(2.29) = 0.011$. If the complexity parameter in `rpart` is reduced below 0.0062, then the CART tree splits on sex. Hsu et al. suggest combining the P-value bounds from the leaves of the tree using Zaykin et al. (2002)'s truncated product of P-values, an extension of Fisher's method of combining P-values. At $\Gamma = 1.64$, if the two P-value bounds for females and males, $1 - \Phi(0.97) = 0.166$ and $1 - \Phi(2.32) = 0.010$, are combined using the truncated product with truncation 0.1, then the combined P-value bound is 0.028. In this one example, the two analyses give fairly similar impressions.

4.4. Simultaneous Inference and Closed Testing

Strictly speaking, the statistic $D_{\Gamma\,\max}$ is a test of a global null hypothesis, specifically Fisher's hypothesis H_0 of no treatment effect in the study as a whole. In previous sections, the c_{gk} are either 0 or 1, and the kth comparison defines a subpopulation $\mathcal{S}_k$ as those groups with $c_{gk} = 1$, that is, $\mathcal{S}_k = \{g : c_{gk} = 1\}$, for instance, the subpopulation of men. We are, of course, interested in the hypothesis, say H_k, that asserts there is no effect in subpopulation $\mathcal{S}_k$, say no effect in the subpopulation of men. We would like to test all K hypotheses H_k, $k = 1, \ldots, K$, strongly controlling the family-wise error rate at α in the presence of a bias of at most Γ. We may do this with the closed testing method of Marcus et al. (1976).

Define $H_{\mathcal{I}}$ for $\mathcal{I} \subseteq \{1, \ldots, K\}$ to be the hypothesis that there is no treatment effect in the union of the subpopulations $\mathcal{S}_k$, $k \in \mathcal{I}$. For instance, in Table 11, the hypothesis $H_{\{2,5\}}$ says that there is no effect for females, $k = 2$, and no effect for smokers, $k = 5$. If $H_{\{2,5\}}$ were true, there might nonetheless be an effect for male nonsmokers. If the goal were to test $H_{\mathcal{I}}$ alone at level α in the presence of a bias of at most Γ, then this could be done using $D_{\Gamma\mathcal{I}} = \max_{k\in\mathcal{I}} D_{\Gamma k}$, which is a test of the same form as $D_{\Gamma\,\max}$, whose approximate critical constant from (4.3.1), say $\kappa_{\Gamma,\alpha,\mathcal{I}}$, must be recalculated using a $|\mathcal{I}|$-dimensional multivariate Normal distribution. Of course, $D_{\Gamma\mathcal{I}} \geq D_{\Gamma\mathcal{J}}$ whenever $\mathcal{J} \subset \mathcal{I}$, so $\kappa_{\Gamma,\alpha,\mathcal{J}} \leq \kappa_{\Gamma,\alpha,\mathcal{I}}$; that is, the correction for multiple testing is less severe when fewer comparisons are made. In particular, $\kappa_{\Gamma,\alpha,\mathcal{I}} \leq \kappa_{\Gamma,\alpha}$ for all $\mathcal{I} \subseteq \{1, \ldots, K\}$.

The closed testing method of Marcus et al. (1976) rejects $H_{\mathcal{I}}$ at level α in the presence of a bias of at most Γ if it rejects $H_{\mathcal{K}}$ for all $\mathcal{K} \supseteq \mathcal{I}$, that is, if $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for all hypotheses $\mathcal{K}$ that contain $\mathcal{I}$. Closed testing has several attractive properties. In general, closed testing strongly controls the family-wise error rate, as demonstrated by Marcus et al. (1976). The extension of this property to sensitivity analyses is straightforward; see Rosenbaum and Silber (2009b), §4.4. That is, no matter which hypotheses are true or false, the probability that closed testing falsely rejects at least one true $H_{\mathcal{I}}$ is at most α whenever the bias is at most Γ. There is an additional property of closed testing that is specific to sensitivity analyses. Use of the Bonferroni inequality in sensitivity analysis is conservative in a way that closed testing is not conservative; see Rosenbaum and Silber (2009b) and Fogarty and Small (2016).

There is a short-cut that simplifies closed testing in this context using the inequality $\kappa_{\Gamma,\alpha,\mathcal{I}} \leq \kappa_{\Gamma,\alpha}$ for all $\mathcal{I} \subseteq \{1,\ldots,K\}$, noted above. Specifically, $D_{\Gamma\mathcal{K}} = \max_{k\in\mathcal{K}} D_{\Gamma k} \geq D_{\Gamma k}$ for all $k \in \mathcal{K}$ and yet $\kappa_{\Gamma,\alpha} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$, so whenever $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha}$ it follows that $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for all hypotheses $\mathcal{K}$ with $k \in \mathcal{K}$. This means that closed testing will reject H_k whenever $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha}$, and may reject H_k in other cases as well. For instance, in Table 11, at $\Gamma = 1.5$, we may reject H_3 and H_7 without calculating $\kappa_{\Gamma,\alpha,\mathcal{K}}$ because $2.77 = D_{\Gamma 3} \geq \kappa_{\Gamma,\alpha} = 2.31$ and $2.45 = D_{\Gamma 7} \geq \kappa_{\Gamma,\alpha} = 2.31$. That is, at $\Gamma = 1.5$, closed testing rejects the null hypothesis of no effect on men and the hypothesis of no effect on the poor.

Consider $\Gamma = 1.4$ in Table 11. The short-cut reject in all groups except females ($k = 2$) and nonpoor ($k = 6$), so that, without further computation, $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for every nonempty $\mathcal{K}$ except $\{2, 6\}$, $\{2\}$, and $\{6\}$. The short-cut does not apply in these cases, so $\kappa_{\Gamma,\alpha,\mathcal{K}}$ must be computed. Using the 2×2 submatrix of $\boldsymbol{\rho}_\Gamma$ for $(D_{\Gamma 2}, D_{\Gamma 6})$, we determine $\kappa_{\Gamma,\alpha,\{2,6\}} = 1.92$, and trivially for $\mathcal{K} = \{2\}$ and $\mathcal{K} = \{6\}$ the critical constant is $\kappa_{\Gamma,\alpha,\mathcal{K}} = 1.64$. Because the short-cut has rejected every $H_{\mathcal{I}}$ with $\{2, 6\} \subset \mathcal{I}$, we compare $D_{\Gamma\{2,6\}} = 2.07$ to $\kappa_{\Gamma,\alpha,\{2,6\}} = 1.92$ and therefore reject $H_{\{2,6\}}$. Continuing, we compare $D_{\Gamma 2} = 1.86$ and $D_{\Gamma 6} = 2.07$ to $\kappa_{\Gamma,\alpha,\mathcal{K}} = 1.64$, and we reject both H_2 and H_6. So, at $\Gamma = 1.40$, some of the $D_{\Gamma k}$ are below $\kappa_{\Gamma,\alpha} = 2.31$, but nonetheless closed testing rejects all seven hypotheses.

It is possible, in principle, to strengthen closed testing when there are logical implications among the hypotheses, $H_1, \ldots, H_K$, as is true here. Here, strengthening means changing the procedure so that it still controls the family-wise error rate but it may, from time to time, reject an additional hypothesis not rejected by closed testing. For instance, Holm (1979) method is the application of closed testing using the Bonferroni inequality, and Shaffer (1986) strengthened Holm's method when applied to the analysis of variance using logical implications among hypotheses. What are the logical implications in Table 11? Recall that the hypotheses assert that no one in certain subpopulations was affected by the treatment. If any of $H_2, \ldots, H_K$ is false, then H_1 is false. Similarly, if H_5 is false, so at least some smokers are affected, then either H_2 or H_3 or both must be false, because every smoker is either male or female. Bergmann and Hommel (1988) discuss the nontrivial general steps required to strengthen a closed testing procedure based on logical implications among hypotheses.

4.5. Aids to Interpreting Subgroup Comparisons

The analysis in §4.4 yields indications of a beneficial effect of physical activity on survival in each subpopulation, but these indications are insensitive to larger biases for men than for women. In the second and third panel of Figure 5, the men are matched for observed covariates, so paired men are similar, as are paired women. However, the men may differ from the women; so, it is useful to examine the observed covariates within subgroups, as is done in Table 14. The men and women are of similar age, but the men are more likely than the women to smoke, drink alcohol, be working, be married, and they have somewhat less education.

The deviates, $D_{\Gamma k}$, in Table 11 may be affected by effect modification, but they are also affected by differing sample sizes. For instance, the deviate for the entire population, $D_{\Gamma 1}$, is based on 470 pairs, whereas the deviate for women, $D_{\Gamma 2}$, is based on 275 pairs of women, and the deviate for men, $D_{\Gamma 3}$, is based on 195 pairs of men. If there were an effect but there were no effect modification — that is, if men and women experienced the same effect of physical inactivity — then we might reasonably expect $D_{\Gamma 1}$ to be larger than $D_{\Gamma 2}$ and $D_{\Gamma 3}$ simply because of the reduced sample size in subpopulations. To separate the sample size and insensitivity to unmeasured bias, a relevant point estimate would be helpful.

It is possible to produce a consistent point estimate of the design sensitivity, $\widetilde{\Gamma}_k$, for the kth statistic. Sample size does not affect the design sensitivity, as it is the limit as the sample size increases without bound. Differing sample sizes alone do not predict an increase or a decrease in the estimated design sensitivity, in contrast with the effect of sample size on the standardized deviates, $D_{\Gamma k}$. This estimate of $\widetilde{\Gamma}_k$ assumes that there is a treatment effect and no unmeasured bias, and then estimates the limiting sensitivity to unmeasured bias as the sample size in this subpopulation increases. In general, $\widetilde{\Gamma}_k$ depends upon the choice of test statistic. In the example, this is the Prentice-Wilcoxon statistic for censored paired survival times, because follow-up ended in 1992 for everyone. Given that the end of follow-up is a fixed date, it is safe to assume that the treatment, physical inactivity, did not affect the length of follow-up. The estimate of $\widetilde{\Gamma}_k$ solves for Γ in the equation $D_{\Gamma k} = 0$ or equivalently in the equation $S_k - \theta_{\Gamma k} = 0$. For all 470 pairs, the estimate of $\widetilde{\Gamma}_1$ is 2.32. For the 275 pairs of women, the estimate of $\widetilde{\Gamma}_2$ is 1.96. For the 195 pairs of men, the estimate of $\widetilde{\Gamma}_3$ is 2.91. In the example, both the deviates $D_{\Gamma k}$ and the estimates of $\widetilde{\Gamma}_k$ suggest there is greater insensitivity to bias for men, and that this is not a consequence of changing sample sizes. In contrast, if the paired survival times for men and for women were independent draws from the same censored bivariate population, then we would expect $D_{\Gamma 2}$ and $D_{\Gamma 3}$ to be smaller than $D_{\Gamma 1}$ because of the reduced sample size, but we would have $\widetilde{\Gamma}_1 = \widetilde{\Gamma}_2 = \widetilde{\Gamma}_3$, so the three point estimates would estimate the same quantity.

4.6. Pairs or Sets that Are Not Exactly Matched for Some Covariates

To avoid confusing a main effect of gender and effect modification involving gender, we search for effect modification in pairs or matched sets that are exactly matched for gender, say in pairs of men, or in pairs of women. In the example in §4.1.1, all pairs were exactly matched for gender, smoking and the indicator of an income above twice the poverty level. With more potential effect modifiers, it may not be possible to match exactly for every

potential effect modifier. What can be done in this case?

If a matched pair were exactly matched for gender, it seems reasonable to use that pair in an analysis that splits on gender, even if the pair is not exactly matched for other potential effect modifiers. Although there may be only a few pairs exactly matched for twenty covariates, it will often be the case that there are many pairs exactly matched for the first covariate, say gender, ignoring the rest, many pairs exactly matched for the second covariate ignoring the rest, and so on. It is straightforward to compare all the pairs of two men, all the pairs of two women, all the pairs of two smokers, etc. It simply requires a small change in the comparison weights, c_{gk}.

Refine the grouping of matched pairs or sets so that there are groups containing only men, groups containing only women, and groups containing matched sets that have both men and women. Then define the comparison weights for men so that a group g of sets containing only men has $c_{gk} = 1$ and all other groups have $c_{gk} = 0$. Define the comparison for women analogously. In this way, there is a comparison for men and a comparison for women, both comparisons use only sets that are exactly matched for gender, some pairs not matched for gender do not get used when analyzing gender, but some of these unused matched sets do get used in other comparisons, say the comparison of smokers.

4.7. Summary and Discussion

4.7.1. Using the submax method to study effect modification and its consequences

Effect modification is important in observational studies for several reasons.

If there were effect modification, then we expect to report firmer causal conclusions in subpopulations with larger effects. That is, we expect the design sensitivity and the power of the sensitivity analysis to be larger, so we expect to report findings that are insensitive to larger unmeasured biases in these subpopulations. Such a discovery is important in three ways. First, the finding about the affected subpopulation is typically important in its own right as a description of that subpopulation. Second, if there is no evidence of an effect in the complementary subpopulation, then that may be news as well. Third, if a sensitivity analysis convinces us that the treatment does indeed cause effects in one subpopulation, then this fact demonstrates the treatment does sometimes cause effects, and it makes it somewhat more plausible that smaller and more sensitive effects in other subpopulations are causal and not spurious. This is analogous to the situation in which we discover that heavy smoking causes lots of lung cancer, and are then more easily convinced that second-hand smoke causes some lung cancer, even though the latter effect is much smaller and more sensitive to unmeasured bias.

Conversely, it can be useful to discover evidence of a treatment effect of the same sign in every major subpopulation. We often worry whether the findings of an observational study in one population can generalize to second population that was not studied. Will a study done in Georgia generalize to Kansas where no study was done? If the second population were simply a different mixture of the same types of people — e.g., in Table 11, a different mixture of men and women, smokers and nonsmokers, rich and poor — then finding strong

evidence of a nontrivial effect of constant sign in all subpopulations gives us some reason to hope that the direction of effect found in the first population will reappear in the second population.

The simulation contrasted the new submax method with another method using groups formed by CART. One big difference between the two methods is that there is more theory concerning the performance of the submax method, including power, design sensitivity and Bahadur efficiency. The submax method achieves the largest design sensitivity of the subgroups, but there is no similar claim for the CART method. In the simulation, the CART method was cautious about forming groups, so it failed to capitalize on slight effect modification, with a loss of power in situation 3; however, that also meant that CART rarely paid a price for multiple testing when there was no effect modification in situation 2. One might tinker with the settings of the CART procedure or the simulation and produce a different result, but that is part of the attraction of the submax method: it has desirable properties that hold in general, without tinkering. In principle, the CART method might discover complex patterns of effect modification that the submax procedure does not consider. More practically, one could combine the two approaches, using the submax procedure with a combination of groups defined a priori, like gender, and a few groups suggested by CART, say poor, nonsmoking, men; however, we have not studied such a joint procedure, in part because it could only be evaluated by simulation.

4.7.2. Other Uses of the Submax Method

Although we have discussed the submax method in §4.3 in the context of effect modification, the same mathematical calculation is useful in other contexts. The method looks at K specified comparisons, $k = 1, \ldots, K$, among the G groups of matched sets using weights $\mathbf{c}_k = (c_{1k}, \ldots, c_{Gk})^T$ of dimension G with $c_{gk} \geq 0$ for $g = 1, \ldots, G$. The $\mathbf{c}_k$ need not pick out subpopulations defined by measured covariates, such as men and women. Two examples will be described briefly. Essentially, the examples distinguish groups of matched sets, but the groups were not formed using the observed covariates, and effect modification is not the concern.

If the treated condition were recorded in G increasing doses or intensities, then we could group matched sets with multiple controls into G groups of sets based on the dose given to the one treated subject in that matched set. The quality or relevance of the dose information may be uncertain. Consider three statistics defined by the comparisons $\mathbf{c}_1 = (1, 1, \ldots, 1)^T$, $\mathbf{c}_2 = (1, 2, \ldots, G)^T$ and $\mathbf{c}_3 = (0, 0, \ldots, 0, 1, 1, \ldots, 1)^T$. The comparison $\mathbf{c}_1$ uses all the matched sets with equal weights, ignoring the doses. The comparison $\mathbf{c}_2$ gives positive weight to all sets, but gives larger weight to sets with higher doses. The comparison $\mathbf{c}_3$ confines attention to sets that received high doses. See Rosenbaum (2010) for calculations of design sensitivities for statistics using doses in different ways. The submax method would use all three statistics, reporting the least sensitive result, adjusting for multiple testing in a manner that reflects the high correlation between three tests that use the same data in different ways.

In an effort to provide information about unmeasured biases, Zubizarreta et al. (2012) produced two types of matched pairs: (i) pairs matched for the hospital providing the

treatment in hospitals that used both the treatment and the control, and (ii) pairs with treated and control patients from different hospitals, one hospital that almost invariably used treatment and another hospital that almost invariably used the control. The first type of pair controls unmeasured covariates that are constant within each hospital, say the hospital's nurse-to-bed ratio. However, in the first type of pair, physicians looked at patients, giving treatment to some patients and control to others, so the first type of pair might be affected by selection bias. In the second type of pair, each patient received the treatment that the hospital almost invariably provides, reducing concern about the selection of individuals for treatment, but the hospitals themselves and the communities they serve may differ in unmeasured ways. In this case, there are $G = 2$ groups of pairs. The comparison $\mathbf{c}_1 = (1, 1)^T$ uses all pairs, $\mathbf{c}_2 = (1, 0)^T$ uses type (i) pairs, and $\mathbf{c}_3 = (0, 1)^T$ uses type (ii) pairs. The submax method would do all three tests with multiple comparisons, as in §4.4, taking account of the high correlation between comparison $\mathbf{c}_1$ and each of the other comparisons.

Table 8: Covariate balance in 470 matched, treatment-control pairs. The standardized difference (Std. Dif) is the difference in means before and after matching in units of the standard deviation before matching.

	Covariate Mean			Std. Dif.	
Covariate	Treated	Control	P-value	Before	After
Age	61.7	61.7	0.985	0.283	0.001
Male	0.415	0.415	1.000	-0.245	0.000
White	0.789	0.823	0.187	-0.252	-0.093
Poverty	0.460	0.460	1.000	0.377	0.000
Former Smoker	0.170	0.145	0.283	-0.142	0.064
Current Smoker	0.360	0.360	1.000	-0.141	0.000
Working last three months	0.247	0.247	1.000	-0.589	0.000
Married	0.621	0.666	0.153	-0.350	-0.099
Dietary Adequacy	3.254	3.379	0.143	-0.303	-0.098
	Education				
≤ 8	0.494	0.466	0.397	0.309	0.057
9-11	0.183	0.204	0.410	-0.097	-0.053
High School	0.166	0.172	0.794	-0.193	-0.016
Some College	0.066	0.070	0.796	-0.158	-0.015
College	0.085	0.085	1.000	0.038	0.000
Missing	0.006	0.002	0.317	0.004	0.054
	Alcohol Consumption				
Never	0.406	0.432	0.428	0.189	-0.053
<1 time per month	0.198	0.185	0.619	0.016	0.032
1-4 times per month	0.172	0.153	0.427	-0.125	0.048
2+ times per week	0.089	0.089	1.000	-0.069	0.000
Just about everyday/everyday	0.134	0.140	0.776	-0.073	0.000

Table 9: Correlation and covariance matrix $\boldsymbol{\rho}_\Gamma$ under H_0 for $D_{\Gamma k}$ for all $\Gamma \geq 1$ in the balanced situation, using Wilcoxon's statistic, with $L = 3$ potential effect modifiers.

	$D_{\Gamma 1}$	$D_{\Gamma 2}$	$D_{\Gamma 3}$	$D_{\Gamma 4}$	$D_{\Gamma 5}$	$D_{\Gamma 6}$	$D_{\Gamma 7}$
$D_{\Gamma 1}$	1.000	0.707	0.707	0.707	0.707	0.707	0.707
$D_{\Gamma 2}$	0.707	1.000	0.000	0.500	0.500	0.500	0.500
$D_{\Gamma 3}$	0.707	0.000	1.000	0.500	0.500	0.500	0.500
$D_{\Gamma 4}$	0.707	0.500	0.500	1.000	0.000	0.500	0.500
$D_{\Gamma 5}$	0.707	0.500	0.500	0.000	1.000	0.500	0.500
$D_{\Gamma 6}$	0.707	0.500	0.500	0.500	0.500	1.000	0.000
$D_{\Gamma 7}$	0.707	0.500	0.500	0.500	0.500	0.000	1.000

Table 10: The critical constant κ_α for $L = 0, \ldots, 15$ balanced binary effect-modifiers, using Wilcoxon's statistic, yielding $K = 2L + 1$ correlated tests with $\alpha = 0.05$. For comparison, the final column gives the critical constant obtained using the Bonferroni inequality, testing K one-sided hypotheses at family-wise level $\alpha = 0.05$.

L	$K = 2L + 1$	κ_α	Bonferroni
0	1	1.64	1.64
1	3	2.03	2.13
2	5	2.20	2.33
3	7	2.32	2.45
4	9	2.40	2.54
5	11	2.46	2.61
6	13	2.51	2.67
7	15	2.55	2.71
8	17	2.59	2.75
9	19	2.62	2.79
10	21	2.65	2.82
11	23	2.67	2.85
12	25	2.70	2.88
13	27	2.72	2.90
14	29	2.74	2.92
15	31	2.75	2.95

Table 11: Seven standardized deviates from Wilcoxon's test, $D_{\Gamma k}$, $k = 1, \ldots, K = 7$, testing the null hypothesis of no effect and their maximum, $D_{\Gamma \max}$, where the critical value is $d_\alpha = 2.31$ for $\alpha = 0.05$. Deviates larger than $d_\alpha = 2.31$ are in **bold**.

k	1	2	3	4	5	6	7	
Subpopulation	All	Female	Male	Non-smoker	Smoker	$> 2\times$ PL	$\leq 2\times$ PL	Maximum
	$D_{\Gamma 1}$	$D_{\Gamma 2}$	$D_{\Gamma 3}$	$D_{\Gamma 4}$	$D_{\Gamma 5}$	$D_{\Gamma 6}$	$D_{\Gamma 7}$	$D_{\Gamma \max}$
Sample-size	470	275	195	301	169	254	216	
$\Gamma = 1.00$	**6.09**	**3.79**	**4.88**	**4.67**	**3.92**	**3.88**	**4.71**	**6.09**
$\Gamma = 1.20$	**4.66**	**2.73**	**3.91**	**3.52**	**3.06**	**2.89**	**3.68**	**4.66**
$\Gamma = 1.40$	**3.48**	1.86	**3.11**	**2.57**	**2.36**	2.07	**2.83**	**3.48**
$\Gamma = 1.60$	**2.47**	1.11	**2.44**	1.76	1.76	1.37	2.10	**2.47**
$\Gamma = 1.64$	2.29	0.97	**2.32**	1.62	1.65	1.24	1.97	**2.32**
$\Gamma = 1.65$	2.24	0.94	2.29	1.58	1.63	1.21	1.94	2.29

Table 12: Theoretical power for Wilcoxon's signed rank test in subgroup analyses using (i) the maximum statistic $D_{\Gamma\,\max}$, (ii) an oracle that knows a priori which group has the largest effect (Oracle), and (iii) one statistic that sums all Wilcoxon statistics, thereby using all the matched pairs, $D_{\Gamma 1}$.

Situation		One covariate, $L = 1$			Five covariates, $L = 5$		
	Γ	$D_{\Gamma\,\max}$	Oracle	$D_{\Gamma 1}$	$D_{\Gamma\,\max}$	Oracle	$D_{\Gamma 1}$
$(\zeta_0, \zeta_1) = (0, 0)$	1	0.050	0.050	0.050	0.050	0.050	0.050
1. No effect. Values	1.01	0.035	0.033	0.033	0.035	0.033	0.033
are the size test.	1.2	0.000	0.000	0.000	0.000	0.000	0.000
	1.3	0.000	0.000	0.000	0.000	0.000	0.000
$(\zeta_0, \zeta_1) = (0.5, 0.5)$	1	1.000	1.000	1.000	1.000	1.000	1.000
2. Constant effect.	2.8	0.579	0.671	0.671	0.460	0.601	0.601
Every subgroup	3.0	0.177	0.218	0.218	0.126	0.167	0.167
has effect 0.5.	3.2	0.030	0.030	0.030	0.020	0.019	0.019
	3.4	0.004	0.002	0.002	0.002	0.001	0.001
$(\zeta_0, \zeta_1) = (0.6, 0.4)$	1	1.000	1.000	1.000	1.000	1.000	1.000
3. Slight effect	2.8	0.991	0.998	0.593	0.959	0.996	0.521
modification,	3.0	0.928	0.971	0.161	0.791	0.959	0.121
$\zeta_0 > \zeta_1$	3.2	0.733	0.855	0.018	0.492	0.816	0.011
	3.4	0.446	0.615	0.001	0.220	0.554	0.001
$(\zeta_0, \zeta_1) = (0.5, 0)$	1	1.000	1.000	1.000	1.000	1.000	1.000
4. Effect confined	2.8	0.268	0.418	0.000	0.113	0.369	0.000
to a subgroup.	3.0	0.071	0.144	0.000	0.020	0.117	0.000
$\zeta_1 = 0$	3.2	0.013	0.033	0.000	0.002	0.025	0.000
	3.4	0.002	0.006	0.000	0.000	0.004	0.000

Table 13: Simulated power (number of rejections in 10,000 replications) for Wilcoxon's signed rank test in subgroup analyses using (i) the maximum statistic $D_{\Gamma\,\max}$, (ii) groups built by CART, (iii) an oracle that knows a priori which group has the largest effect (Oracle), and (iv) one statistic that sums all of the Wilcoxon statistics, thereby using all matched pairs, $D_{\Gamma 1}$.

		One covariate, $L = 1$				Five covariates, $L = 5$			
$\mu = (\mu_0, \mu_1)$	Γ	$D_{\Gamma\,\max}$	CART	Oracle	$D_{\Gamma 1}$	$D_{\Gamma\,\max}$	CART	Oracle	$D_{\Gamma 1}$
(0,0)	1	540	525	525	525	515	504	503	503
	1.01	382	344	344	344	345	329	328	328
	1.1	7	1	1	1	7	7	7	7
	1.2	0	0	0	0	1	0	0	0
	1.3	0	0	0	0	0	0	0	0
(0.5, 0.5)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	5804	6713	6713	6713	4581	6014	6014	6014
	3.0	1643	2104	2101	2101	1215	1685	1681	1681
	3.2	279	315	313	313	158	187	183	183
	3.4	30	13	12	12	11	10	9	9
(0.6, 0.4)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	9913	7073	9977	6058	9589	6584	9955	5348
	3.0	9264	3788	9701	1657	7975	3471	9588	1242
	3.2	7387	2313	8565	173	5071	2212	8208	121
	3.4	4603	1535	6265	6	2245	1363	5679	8
(0.5, 0)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	2687	1908	4195	0	1105	1626	3686	0
	3.0	678	514	1476	0	174	398	1139	0
	3.2	120	100	320	0	23	67	208	0
	3.4	16	14	47	0	3	10	31	0

Table 14: Covariate means in 275 pairs of women and 195 pairs of men.

	Covariate Mean			
	Female		Male	
	Treated	Control	Treated	Control
Sample size	275	275	195	195
Age	61.2	61.0	62.5	62.7
Male	0.000	0.000	1.000	1.000
White	0.775	0.822	0.810	0.826
Education				
0-8	0.476	0.429	0.518	0.518
9-11	0.211	0.211	0.144	0.195
High School	0.185	0.207	0.138	0.123
Some College	0.069	0.084	0.062	0.051
College	0.051	0.069	0.133	0.108
Missing	0.007	0.000	0.005	0.005
Poverty	0.476	0.476	0.436	0.436
Former Smoker	0.116	0.080	0.246	0.236
Current Smoker	0.273	0.273	0.482	0.482
Working last three months	0.193	0.189	0.323	0.328
Married	0.502	0.553	0.790	0.826
Dietary Adequacy	3.045	3.139	3.549	3.716
Alcohol Consumption				
<1 time per month	0.222	0.222	0.164	0.133
1-4 times per month	0.116	0.135	0.251	0.179
2+ times per week	0.051	0.069	0.144	0.118
Just about everyday/everyday	0.084	0.084	0.205	0.221
Never	0.527	0.491	0.236	0.349

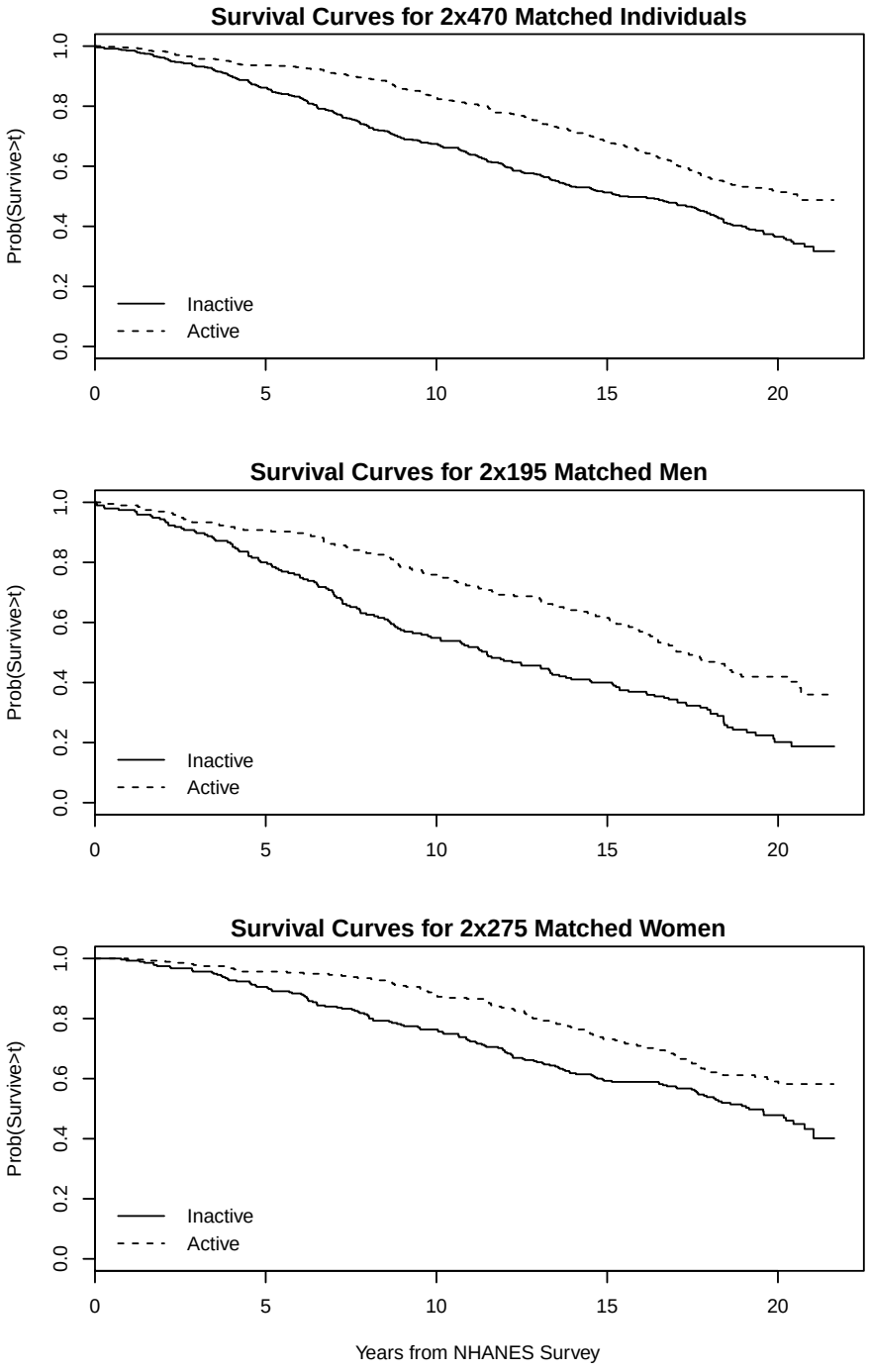

Figure 5: Survival in inactive and matched active groups following the NHANES survey.

CHAPTER 5 : Estimating the Malaria Attributable Fever Fraction Accounting for Parasites Being Killed by Fever and Measurement Error

5.1. Introduction

Malaria is a mosquito-borne infectious disease caused by a parasite. In many tropical regions, malaria is a giant-killer of children, imposes financial hardship on poor households, and holds back economic growth and improvements in living standards (World Health Organization, 2011). The most characteristic clinical feature of malaria is fever (Warrell, 1993). The malaria attributable fever fraction (MAFF) for a group of people is the proportion of fevers in the group of people that are attributable to (caused by) malaria. We will consider the MAFF for children in sub-Saharan Africa, the population hit hardest by malaria (World Health Organization, 2011). The MAFF is an important public health quantity for several reasons that include

- The MAFF provides information about the public health burden from malaria and how much resources should be devoted to combatting malaria compared to other diseases (Mabunda et al., 2009).

- The MAFF is an essential input to the prevalence of malaria attributable fevers (PMAF); the PMAF equals MAFF $\times$ prevalence of fevers. Changes in the PMAF over time provide information about the effects of public health programs that combat malaria (Koram and Molyneux, 2007).

- For planning the sample size for a clinical trial of an intervention against malaria, the MAFF is an essential input (Halloran et al., 1999; Smith, 2007). For example, suppose we are planning a trial of duration one year and want to have 80% power for an intervention that halves malaria attributable fever but has no effect on other sources of fever, and in the population of interest, the average child suffers from 10 fevers per year. The needed sample size depends on the MAFF and is about $n = (800, 1325, 2500)$ in this example if the MAFF is $(0.5, 0.4, 0.3)$ respectively.

- For clinicians treating a child suffering from fever and needing to decide how to prioritize providing antimalarial treatment vs. treatments for other possible sources of the fever, knowing the MAFF conditional on the child's symptoms (e.g., the intensity of the fever and the child's parasite density) is a valuable input (Koram and Molyneux, 2007). In particular, letting 'MAFF | Symptoms' denote the MAFF conditional on the child's symptoms, a doctor would want to treat a patient with an anti-malarial if (Expected gain in utility from treating child at time t with an anti-malarial if she has a malaria attributable fever vs. not treating)$\times$(MAFF | Symptoms) $>$ (Expected loss in utility from treating child at time t with an anti-malarial if she has a fever that is not malaria attributable fever vs. not treating) $\times$ [1-(MAFF | Symptoms)].

The MAFF could be estimated from a survey by a usual ratio estimator if it was easy to determine whether or not a fever was attributable to malaria parasites. However, fevers caused by malaria parasites often cannot be distinguished on the basis of clinical features from fevers caused by other common childhood infections such as the common cold, pneu-

monia, influenza, viral hepatitis or typhoid fever (Hommel, 2002; Koram and Molyneux, 2007). One aid to deciding whether a fever is caused by malaria parasites or some other infection is to measure the density of malaria parasites in the child's blood. However, in areas where malaria is highly endemic, children can develop partial immunity to the toxic effects of the parasites and can tolerate high parasite densities without developing fever (Marsh, 2002; Boutlis et al., 2006). Consequently, even if a child has a fever and has a high parasite density, the fever might still be caused by another infection. In summary, it cannot be determined with certainty whether a given child's fever is malaria attributable, making estimating the MAFF challenging.

In this paper, we make two contributions to estimating the MAFF. First, we provide an analysis of the assumptions needed for existing estimators of the MAFF to be consistent and show that these assumptions are not plausible. Although there has been numerous previous work on estimating the MAFF (Greenwood et al. (1987), Smith et al. (1994), Vounatsou et al. (1998), Qin and Leung (2005), Wang and Small (2012)), estimators have been proposed without clearly defining the estimand. We use the potential outcomes framework to clearly define the estimand and make clear the causal assumptions on which the consistency of these previous estimators rest. These assumptions include that non-malaria attributable fevers do not kill parasites and that there is no measurement error of a certain type in measuring parasite density. We discuss evidence that these assumptions are *not* plausible in most settings, and show that existing MAFF estimation methods are biased under plausible violations of the assumptions.

The second major contribution of our paper is that we develop a consistent estimator of the MAFF that allows for parasites being killed by fever and measurement error in parasite density. Our novel estimation method extends the g-modeling method for solving deconvolution problems (Efron, 2016) to the setting of malaria survey data, accounting for parasites being killed by fever and measurement error. Specifically, survey data on malaria can be divided into two groups, the children with fever (the febrile group) and the children without fever (the afebrile group). The group with fever is a mixture of two components: children with a fever that is malaria attributable and children with a fever that is not malaria attributable. The group without fever can be used as a training sample that provides information about the distribution of parasite densities of the latter mixture component. The main idea of our method is to recover the distributions of the parasite densities of the mixture components before fever killing and measurement error by assuming that the mixture components are in flexible exponential families and solving the deconvolution problem. Using simulation studies, we show that our proposed method produces approximately unbiased estimates of the MAFF when the magnitude of fever killing and the measurement error mechanism are correctly specified. We apply our method to make inferences about the MAFF for a study area in Kilombero, Tanzania.

The rest of this article is organized as follows. In Section 5.2, we define the MAFF and state critical assumptions in the potential outcome framework. We also reveal the relationship between the potential outcome framework MAFF and the existing estimators of the MAFF based on observable quantities under the assumptions. Also, we describes possible violations of the assumptions that there is no fever killing and no measurement error. Also, the

impact of a violation of these assumptions on current estimation methods is investigated. In Section 5.3, we develop our new estimation method, the maximum likelihood estimation method using the g-modeling approach. Section 5.4 shows the performance of our proposed method with simulation studies and Section 5.5 shows the application to the malaria data in Tanzania with a sensitivity analysis. Section 5.6 includes summary and discussion.

5.2. Malaria Attributable Fever Fraction

5.2.1. Potential Outcome Definition of the MAFF

To define a fever as being caused by (i.e., attributable to) malaria parasites, we use the Neyman-Rubin potential outcomes framework for causal inference (Neyman, 1923, 1990). For each child i and each possible parasite density level d, the potential outcome $Y_i^{(d)}$ is 1 or 0 according to whether or not the child would have fever if an intervention set the child's parasite density level to d (the intervention does not need to be specified but need to satisfy Assumption 2 which we will discuss below; for example, an intervention can be malaria prevention program such as antimalarial drugs, vaccines and control of the mosquitoes that carry the malaria parasites). Each child has many potential outcomes, but we observe only one potential outcome, $Y_i^{obs} \equiv Y_i^{(D_i)}$, where D_i is the child's actual parasite density (we are only able to measure D_i with some error; see Section 5.2.2). In addition to the actually observed outcome Y_i^{obs}, we consider two latent variables, called Y_i^{nmi} and Y_i^{mi}, that represent a fever caused by non-malarial infections and a fever caused by malarial infections respectively. The observed fever Y_i^{obs} is represented as $Y_i^{obs} = \max\{Y_i^{nmi}, Y_i^{mi}\}$, which means that if a fever is observed, either non-malarial infections or malarial infections or both triggered the fever. That is, the way malaria and non-malaria infections affect a fever is like parallel circuits. Interestingly, Y_i^{nmi} can be understood by using the potential outcome $Y_i^{(0)}$ that is the outcome under an intervention that eliminates malaria parasites from child i's body. In the absence of malaria parasite in blood, the only possible source of fever were non-malarial infections. Therefore, when child i had a non-malarial fever (i.e, $Y_i^{nmi} = 1$), if an intervention eliminated all parasites in blood, then the fever would remain (i.e., $Y_i^{(0)} = 1$). This implies $Y_i^{nmi} = Y_i^{(0)}$, so we will use this property for the rest of this article.

We define a child i who is observed to have a fever as having a *malaria attributable fever* if that fever would not have occurred if the child had been given an intervention that prevented the child from having malaria parasites. In terms of potential outcomes, a child i has a malaria attributable fever if $Y_i^{(D_i)} = 1$, but $Y_i^{(0)} = 0$ or alternatively, if $Y_i^{obs} = 1$ but $Y_i^{nmi} = 0$. Let (Y^{mi}, Y^{nmi}) be the random vector from the experiment of choosing a random child and time point from the study area and study period. The fraction of fevers in the study area and time period that are attributable to malaria, i.e., the MAFF, is

$$MAFF = \Pr(Y^{(0)} = 0 | Y^{(D)} = 1) = \Pr(Y^{nmi} = 0 | Y^{obs} = 1). \tag{5.2.1}$$

Because we have defined the MAFF using potential outcomes based on an intervention that could alter parasite density, the MAFF could depend on the intervention that alters

parasite density, e.g., possible interventions are antimalarial drugs, vaccines and control of the mosquitoes that carry the malaria parasites. Consider an intervention that satisfies:
Assumption 2. *No Side Effects Assumption.* The intervention has no effects on fever beyond removing the parasites from the child.

Assumption 2 implies that the intervention cannot cause a fever:

$$P(Y^{nmi} = 1, Y^{obs} = 0) = 0. \tag{5.2.2}$$

Under Assumption 2, the MAFF can be interpreted as the proportion of fevers in the observed world that would be eliminated if malaria were eradicated by the intervention. We will not specify the intervention under consideration but will assume that the intervention satisfies Assumption 2. Although it is possible that interventions currently being studied could have side effects and violate Assumption 2, the MAFF for a hypothetical intervention that satisfies Assumption 2 provides an estimate of the *potential* benefit of an intervention to eliminate malaria which is useful for policymakers (Walter, 1976).

The potential outcome framework MAFF (5.2.1) is defined based on unobservable latent variables Y^{nmi} and Y^{mi}. In order to identify the MAFF from data, assumptions need to be made about the way these unobserved variables link to observed variables. To estimate the MAFF, we make the following assumptions:
Assumption 3. The potential outcome $Y^{(d)}$ satisfies that

(i) (Monotonicity Assumption) For any $0 \leq d \leq d'$, $Y^{(d)} \leq Y^{(d')}$

(ii) A fever caused by a non-malaria infection Y^{nmi} is independent of a fever caused by a malaria infection Y^{mi}, i.e. $Y^{nmi} \perp\!\!\!\perp Y^{mi}$.

Assumption 3 (i), the monotonicity assumption, is biologically plausible because having more parasites means that more red blood cells are ruptured by the parasites and more hemozoin is released, meaning that if a child's hemozoin level was enough to cause a fever at parasite level d, then a child would surely have a fever at parasite level d' since the child's hemozoin level would be even higher. Assumption 3 (ii) means that two causes of a fever, non-malarial infections and malarial infections, act separately to trigger a fever. However, this assumption might not hold if Y^{nmi} and Y^{mi} have an impact on each other. For example, imagine someone has a severe cold, he or she have weaker immune system and is more likely to be vulnerable to malarial infections. These malarial infections may cause a fever that would not be triggered if he or she did not have a cold, which implies Y^{nmi} and Y^{mi} can be positively correlated. Also, the assumption can be violated if Y^{nmi} and Y^{mi} have some confounding variables, such as age or health condition. For the former situation, the impact of the violation can be analyzed by conducting sensitivity analyses discussed in Section 5.3.3. The latter violation can be eased by controlling for other observed covariates. We will discuss incorporting covariates into our estimation method in Section 5.3.1.

The conventional approaches for estimating the MAFF have been designed to estimate associative measures of the MAFF that do not entail causal interpretation. Let p_f be the parasite prevalence in febrile children and p_a be the parasite prevalence in afebrile children.

That is, $p_f = \Pr(D^{obs} > 0 | Y^{obs} = 1)$ and $p_a = P(D^{obs} > 0 | Y^{obs} = 0)$ where D^{obs} is the observed parasite density. One popular measure is based on relative risk (Smith et al., 1994) which is defined as $p_f(R-1)/R$ where $R = \Pr(Y^{obs} = 1 | D^{obs} > 0) / \Pr(Y^{obs} = 1 | D^{obs} = 0)$ is the relative risk of fever associated with the exposure of parasites. Greenwood et al. (1987) proposed a simple estimator that p_f and R are directly estimated from the observed data. For more complicated methods on estimating the relative risk measure, see Smith et al. (1994); Wang and Small (2012). Another popular associative measure is based on odds ratio, i.e., $OR = p_f(1-p_a)/(p_a(1-p_f))$. The relative risk is often approximated by the odds ratio, so the relative risk can be replaced by the odds ratio. The measure on odds ratio is defined as $(p_f - p_a)/(1-p_a)$. Many sophisticated methods have been developed to estimate this measure (Vounatsou et al., 1998; Qin and Leung, 2005). The existing methods using the relative risk measure actually estimate the same quantity as the MAFF defined by potential outcome if Assumption 1 and 2 hold and, additionally, the density is correctly measured. On the other hand, the existing methods using the odds ratio measure need an additional step for adjustment to estimate the potential outcome MAFF under the assumptions. If the assumptions are violated, then all the existing methods would provide biased estimates. We discuss the connection between the associative measures and the potential outcome measure in Appendix with more detail.

5.2.2. Fever killing effect and measurement error

To investigate the MAFF, it is required to observe parasite density D as well as the outcome of fever. However, parasite density can be observed with some errors because of fever killing and measurement error. Fever killing refers to the fact that a fever kills some parasites in the body and measurement error refers to the fact that it is difficult to measure parasite density with great accuracy. We provide a brief review of fever killing and measurement error, and provide a model describing them.

To account for fever killing and measurement error, we consider three different variables related to the parasite density: $D_i^{no.nmi}, D_i^{cur}$ and D_i^{obs}. Let $D_i^{no.nmi}$ be the parasite density that a subject i would have if the subject does not have a non-malaria infection strong enough to cause a fever (Small et al., 2010). This is the true parasite density if there is no fever killing and no error in measuring malaria parasites. Therefore, we ultimately want to have $D_i^{no.nmi}$ for analysis. If fever killing occurs, then the parasite density may be changed. We denote D_i^{cur} as the amount of the parasite density in blood of the subject after fever killing occurs. If a subject i has a fever that is solely caused by a non-malaria infection, $Y_i^{nmi} = 1, Y_i^{mi} = 0$, then there is evidence that fever kills some of the parasites that would have remained alive in the absence of the infection (Kwiatkowski, 1989; Rooth and Bjorkman, 1992; Long et al., 2001). In particular, Long et al. (2001) estimate that a fever of 38.8°C kills 50 % of parasites and a fever of 40°C kills 92 % of parasites. Fever killing will make $D_i^{no.nmi}$ greater than the actual current parasite density. See Small et al. (2010) for more discussion on fever killing.

Fever killing occurs in some sense for all fevers, however we define $D_i^{no.nmi}$ for malarial fevers ($Y_i^{mi} = 1$) in such a way that fever killing in terms of D_i^{cur} being different from $D_i^{no.nmi}$ occurs only for non-malarial fevers $Y_i^{nmi} = 1, Y_i^{mi} = 0$ Specifically, we define $D_i^{no.nmi}$ in the

following way for a child with a malarial fever, $Y_i^{mi} = 1$. In a malaria infection that is not brought under control by a child's immunity, the parasites multiply inside the red blood cells they invade, eventually causing the red blood cell to rupture, and the released parasites then invade new red blood cells (Kitchen, 1949). This causes an exponential growth phase of the parasites that terminates shortly after the onset of periodic fever (Kitchen, 1949). The fever starts killing parasites while at the same time the parasites that remain alive continue to multiply. The clash of these two forces and the features of the parasite life cycle cause the parasite density to oscillate (Kwiatkowski and Nowak, 1991). For a child with a malaria fever $Y_i^{mi} = 1$, even if any non-malarial infection were removed, this process of the parasites multiplying and eventually rising above the pyrogenic threshold and then oscillating would occur. For a child with a malarial fever, we define the child's parasite density $D_i^{no.nmi} = D_i^{cur}$ as the parasite density at the point in the child's fever at which the child is observed.

Besides fever killing, measurement error occurs when researchers measure the actual current parasite density D_i^{cur}. Let D_i^{obs} be the observed (measured) parasite density. For example, since it is impossible to count all malaria parasites in blood, doctors typically take a blood sample of an individual and estimate the total amount of the parasites in blood from the sample. During this procedure, some measurement errors may occur, and therefore we only observe (or measure) D_i^{obs} with some errors. Dowling and Shute (1966) and O'Meara et al. (2007) study the sources and magnitude of measurement error. The sources include the following: (1) Sample variability. The parasite density is estimated from a sample of blood; (2) Loss of parasites in the sample handling and staining process; (3) Microscopy error. The accuracy of the parasite density measurement depends on the quality of the microscope and the concentration and motivation of the microscopist; (4) Sequestration and synchronization. Microscopic examination of a blood sample only estimates the parasite density in the peripheral blood and not the total parasite density. Older parasites sequester in the vascular beds of organs. Due to a tendency of the life cycles of the parasites to be synchronized, there can be large variation in the parasite density in the peripheral blood relative to the total parasite density (Bouvier et al., 1997); (5) Variability in white blood cell density. The most common method of estimating parasite density counts the number of parasites found for a fixed number of white blood cells and then assumes that there are 8000 white blood cells per μl of blood. White blood cell counts actually vary considerably from person to person and from time to time within a person (McKenzie et al., 2005).

Figure 1 summarizes the relationship between all the defined variables. As can be seen, $D_i^{no.nmi}$ is the only parasite density variable that decides whether child i has a malaria fever or not while D_i^{cur} and D_i^{obs} are proxy variables of $D_i^{no.nmi}$. We can think of $D_i^{no.nmi}$ as the parasitological challenge faced by the child which is a function of the amount of the parasites injected from mosquito bites and the immune response of the child. The other two parasite density variables, D_i^{cur} and D_i^{obs}, change according to both fever killing and measurement error but do not directly affect whether the child has a malarial fever.

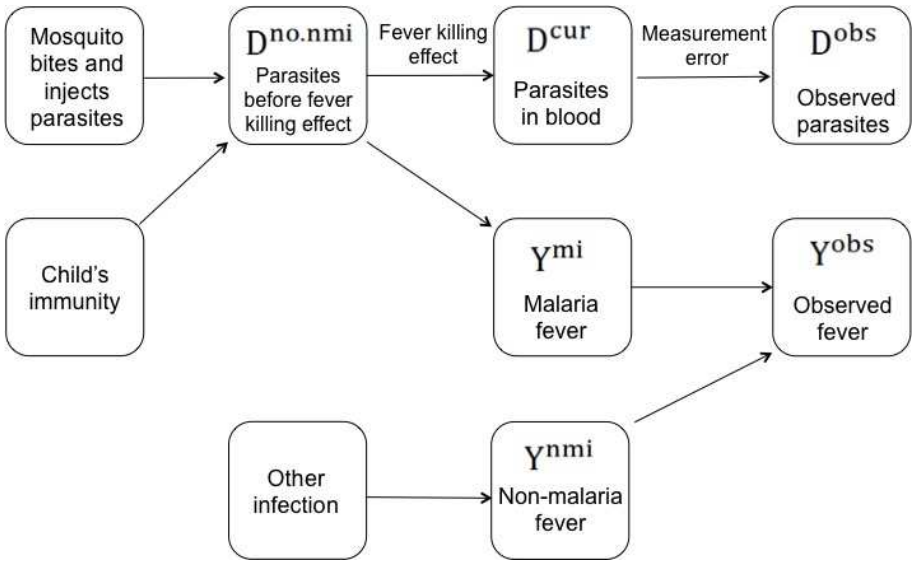

Figure 6: Causal diagram

5.3. Estimation Method

There are n individuals. Each individual has two observable variables and a vector of covariates $(Y_i^{obs}, D_i^{obs}, \mathbf{X}_i)$. The observed fever outcome Y_i^{obs} is binary, but D_i^{obs} is usually continuous. In the Kilombero study that we will discuss later, $\mathbf{X}_i$ contains intercept and age, but we do not consider this covariate vector in this section. $\mathbf{X}_i$ is not required in our estimation method if the independent assumption of Y^{nmi} and Y^{mi} holds, but controlling for $\mathbf{X}_i$ can make the assumption more plausible. Incorporating $\mathbf{X}_i$ will be discussed in Section 5.3.4. In our discussion of methods for estimating the MAFF, we will assume that we have a random sample such that the random vectors $(Y_i^{obs}, D_i^{obs}, \mathbf{X}_i, Y_i^{nmi}, Y_i^{mi}, D_i^{cur}, D_i^{no.nmi}), i = 1, \ldots, n$ are i.i.d. with the same distribution as the random vector $(Y^{obs}, D^{obs}, \mathbf{X}, Y^{nmi}, Y^{mi}, D^{cur}, D^{no.nmi})$.

5.3.1. Bayes deconvolution problem and g-modeling

As shown in Figure 1, Y^{obs} and D^{obs} are associated due to the true parasite density $D^{no.nmi}$. Therefore, we build a model for the true density $D^{no.nmi}$ instead of D^{obs}. To model this density, we use parametric exponential family distribution modelling, called *g-modeling*, proposed by Efron (2016). The g-modeling is the approach to estimate the density of a latent variable such as $D^{no.nmi}$ from the observable variable in Bayes deconvolution problems. In the malaria study, the deconvolution process is both fever killing effects and measurement error. Efron shows that parametric exponential family modeling can give useful estimates in moderate-sized samples while traditional asymptotic calculations are discouraging, indicating very slow nonparametric rates of convergence. Although this approach assumes a parametric model, it can be made more flexible by increasing the number of parameters as the sample size increases. The problem of overfitting from using too many parameters can be avoided by penalizing the likelihood. With enough parameters, the risk of model misspecification is reduced.

To model the fever killing effect, we use a parameter β for the size of the effect; if $\beta = 0.1$, then the non-malaria caused fever kills 90% of parasites. Fever killing occurs only when $Y^{nmi} = 1, Y^{mi} = 0$. Therefore, D^{cur} is equal to $\beta D^{no.nmi}$ if $Y^{nmi} = 1, Y^{mi} = 0$, and D^{cur} is the same as $D^{no.nmi}$ otherwise. For measurement error, let h be the measurement error mechanism. D^{obs} has the density function with the parameter D^{cur}, i.e. $D^{obs} = x|D^{cur} \sim h(x; D^{cur})$. For example, one way to measure malaria parasites is to count the number of parasites in a fixed volume of sampled blood (Earle et al., 1932); this measurement method was advocated by McKenzie et al. (2005). If this measurement error method is used and the only source of measurement error is the sampling of blood, then $D^{obs}|D^{cur} \sim Poisson(D^{cur})$. We assume that the size of the fever killing effect and the measurement error mechanism are known to researchers; if these are unknown, they can be varied in a sensitivity analysis. The assumptions of a known $100(1-\beta)\%$ fixed size of fever killing and known measurement error mechanism h allow us to estimate the MAFF through a Bayes deconvolution approach under Assumptions 2 and 3.

Define

$$g_1(z) \equiv f(D^{no.nmi} = z|Y^{mi} = 0)$$
$$g_2(z) \equiv f(D^{no.nmi} = z|Y^{mi} = 1).$$

Since Y^{nmi} and $D^{no.nmi}$ are independent in Assumption 2, the conditional density $f(D^{no.nmi} = z|Y^{mi})$ is independent of Y^{nmi}. This implies that $f(D^{no.nmi} = z|Y^{mi} = 0) = f(D^{no.nmi} = z|Y^{mi} = 0, Y^{nmi} = 0) = f(D^{no.nmi} = z|Y^{obs} = 0)$. Similarly, we can construct the conditional density $f(D^{no.nmi} = z|Y^{obs} = 1)$ as a mixture of two densities,

$$f(D^{no.nmi} = z|Y^{obs} = 1) = (1 - \lambda^*)g_1(z) + \lambda^* g_2(z),$$

where $\lambda^* = \Pr(Y^{mi} = 1)/\Pr(Y^{obs} = 1)$.

To apply the g-modeling approach, we assume that the two densities $g_1(z)$ and $g_2(z)$ have the form of exponential family distributions. Since the parasite density is non-negative (i.e., $g_1(z) = 0$ and $g_2(z) = 0$ for $z < 0$) and no parasite cannot cause malarial infections (i.e., $g_2(0) = 0$), we give a specific form of the model. We also consider an exponential tilt model between $g_1(z)$ and $g_2(z)$ when $z > 0$, we assume the following model,

$$\begin{aligned} g_1(z; q, \alpha) &= q \cdot I(z = 0) + (1 - q) \cdot \exp\{Q_z^T \alpha - \phi_1(\alpha)\} I(z > 0), \quad 0 \le q \le 1 \\ g_2(z; \alpha, \gamma) &= \exp\{\gamma_0 + \gamma_1 z\} \cdot \exp\{Q_z^T \alpha - \phi(\alpha)\} = \exp\{Q_z^T \alpha + \gamma z - \phi_2(\alpha, \gamma)\} \end{aligned} \tag{5.3.1}$$

where α is a m-dimensional parameter, Q_z is a smoothly defined $m \times 1$ vector function of x, and $\phi_1(\alpha) = \log\{\int \exp(Q_z^T \alpha) dz\}$. Similarly, $\phi_2(\alpha, \gamma) = \log\{\int \exp(Q_z^T \alpha + \gamma z) dz\}$. For further computational details, see Efron (2016).

5.3.2. Estimation

Given the density functions $g_1(z)$ and $g_2(z)$ defined in the previous section, we can describe the observed data. The overall observed data can be partitioned into two groups: a febrile group that contains individuals with $Y^{obs} = 1$ and an afebrile group with $Y^{obs} = 0$. For

simplicity, let the sample for the afebrile group be $\mathbf{D}_0^{obs} := (D_{01}^{obs}, \ldots, D_{0n_0}^{obs})$ and the sample for the febrile group be $\mathbf{D}_1^{obs} := (D_{11}^{obs}, \ldots, D_{1n_1}^{obs})$ with $n = n_0 + n_1$. Then, the two samples are drawn from the following observed parasite densities,

$$D_{01}^{obs}, \ldots, D_{0n_0}^{obs} \sim \int h(d; z) \cdot g_1(z; q, \alpha) dz$$
$$D_{11}^{obs}, \ldots, D_{1n_1}^{obs} \sim \int h(d; z) \cdot \{(1 - \lambda^*) g_1^*(z; q, \alpha) + \lambda^* g_2(z; \alpha, \gamma)\} dz$$

where $g_1^*(z) = \beta^{-1} g_1(z/\beta)$ that incorporate fever killing effect and h is the known measurement error mechanism.

Denote $p := \Pr(Y^{obs} = 1)$. Then, the likelihood $\mathcal{L}(\mathbf{D}_0^{obs}, \mathbf{D}_1^{obs}; p, \lambda^*, q, \alpha, \gamma)$ is written as

$$\begin{aligned}
&\mathcal{L}(\mathbf{D}_0^{obs}, \mathbf{D}_1^{obs}; p, \lambda^*, q, \alpha, \gamma) \\
=& \Pr(Y^{obs} = 0)^{n_0} \Pr(Y^{obs} = 1)^{n_1} \prod_{i=1}^{n_0} P(D^{obs} = D_{0i}^{obs} | Y^{obs} = 0)) \prod_{i=1}^{n_1} P(D^{obs} = D_{1i}^{obs} | Y^{obs} = 1) \\
=& (1-p)^{n_0} p^{n_1} \times \prod_{i=1}^{n_0} \left\{ \int h(D_{0i}^{obs}; z) \cdot g_1(z; q, \alpha) dz \right\} \\
& \times \prod_{i=1}^{n_1} \left\{ \int h(D_{1i}^{obs}; z) \cdot \{(1 - \lambda^*) g_1^*(z; q, \alpha) + \lambda^* g_2(z; \alpha, \gamma)\} dz \right\}.
\end{aligned}$$

The log-likelihood $\ell(\mathbf{D}_0^{obs}, \mathbf{D}_1^{obs}; p, \lambda^*, q, \alpha, \gamma)$ is written as the sum of two parts, $\ell_1(p) + \ell_2(\lambda^*, q, \alpha, \gamma)$, where

$$\begin{aligned}
\ell_1(p) &= n_0 \log(1 - p) + n_1 \log p \\
\ell_2(\lambda^*, q, \alpha, \gamma) &= \sum_{i=1}^{n_0} \log \left\{ \int h(D_{0i}^{obs}; z) \cdot g_1(z; q, \alpha) dz \right\} \\
&+ \sum_{i=1}^{n_1} \log \left\{ (1 - \lambda^*) \cdot \int h(D_{1i}^{obs}; z) \cdot g_1^*(z; q, \alpha) dz + \lambda^* \cdot \int h(D_{1i}^{obs}; z) g_2(z; \alpha, \gamma) dz \right\}.
\end{aligned} \tag{5.3.2}$$

The estimates $\hat{p}$ can be obtained by maximizing the log-likelihood $\ell_1 p$ and $\hat{\lambda}^*$ can be obtained by maximizing the log-likelihood $\ell_2(\lambda^*, q, \alpha, \gamma)$. Then, the estimate $\hat{\lambda}$ of the MAFF is obtained from $\hat{\lambda}^*$ and $\hat{p}$ by using adjustment $\lambda = \lambda^*(1 - p)/(1 - p\lambda^*)$.

To apply the g-modeling approach, a practical question that can be raised is the choice of the dimension of the parameter vector α. An increase of the dimension of α would provide a higher likelihood value, but excessive number of parameters may cause problems such as overfitting. To avoid the overfitting problem, we choose a model based on Bayesian information crierian (BIC). Haughton et al. (1988) discussed this issue for exponential family distributions, and analytically proved choosing a model based on BIC leads to a correct choice of a model with high probability. An alternative is cross-validation approaches to find an optimal dimension of α using likelihood-based cross-validation approaches (van der

Laan et al., 2004). For adequate smoothness of the exponential family distribution fitting, we suggest using either BIC or cross-validation approaches to choose the dimension of α.

Efron (2016) suggests to use a penalized likelihood as he finds that the accuracy of a deconvolution estimate obtained from the g-modeling approach can be greatly improved by regularization of the maximum likelihood algorithm. Instead of maximizing $\ell(p, \lambda^*, q, \alpha, \gamma)$, we maximize a penalized log-likelihood

$$m(p, \lambda^*, q, \alpha, \gamma) = \ell(p, \lambda^*, q, \alpha, \gamma) - s(\alpha)$$

where $s(\alpha)$ is a penalty function. In this article, the function $s(\alpha)$ is defined as $s(\alpha) = c_0\|\alpha\|$ where c_0 is a regularizing constant. The choice of c_0 is related to bias-variance tradeoff. Efron (2016) discussed evaluation of the choice of c_0, and we take a similar evaluation strategy of choosing c_0 specific to our likelihood. In the Kilombero study data, $c_0 = 50$ is as a modest choice that restricts the trace of the added variance due to the penalization, see Appendix A.4.1. For general computation details, see Efron (2016), Remark A4.

5.3.3. Sensitivity Analysis for Assumption 3

Our proposed estimation method relies on Assumption 3, particularly, Assumption 3 (ii) that assume Y^{mi} and Y^{nmi} have independent causal pathways that form a parallel circuit to trigger a fever. However, these assumptions could potentially be violated. For example, consider a situation such that a child has some malaria parasites in blood, but it is not strong enough to trigger a malarial fever. Without any further malarial infections, if a child got a cold, this non-malarial infection might trigger a malarial fever because a combined effect of some malaria parasites and some non-malarial infections might weaken the child's immune system to be enough to trigger a fever. We consider this

If the independent assumption is violated, the following equalities do not hold: (1) $f(D^{no.nmi} = z|Y^{nmi} = 0, Y^{mi} = 0) = f(D^{no.nmi} = z|Y^{nmi} = 1, Y^{mi} = 0)$ and (2) $\Pr(Y^{nmi} = 0, Y^{mi} = 0) = \Pr(Y^{nmi} = 0)\Pr(Y^{mi} = 0)$. If the two equal relationships are broken, then it is impossible to identify the MAFF from the data. Instead, in this situation, we can consider sensitivity parameters to describe and restrict the relationships. By doing so, the range of the MAFF can be obtained. For describing the former relationship, an exponential till model can be used as

$$f(D^{no.nmi} = z|Y^{nmi} = 1, Y^{mi} = 0) = \exp(\delta_0 + \delta_1 z) f(D^{no.nmi} = z|Y^{nmi} = 0, Y^{mi} = 0).$$

The parameter δ_0 is identified because of $\int f(D^{no.nmi} = z|Y^{nmi} = 1, Y^{mi} = 0)\, dz = 1$, and is a function of δ_1. Similar to the estimation method in Section 5.3, we use the exponential family distribution model for the density $f(D^{no.nmi} = z|Y^{nmi} = 0, Y^{mi} = 0) = \exp\{Q_z\alpha - \phi_1(\alpha)\}$ where $\phi_1(\alpha) = \log\left(\int \exp\{Q_z\alpha\}\right)$. Therefore, the parasite density $f(D_i^{no.nmi} = d_j|Y_i^{mi,*} = 0, Y_i^{nmi} = 1)$ is represented as

$$\exp(\delta_0 + \delta_1 z)\exp\{Q_z\alpha - \phi_1(\alpha)\} = \exp\{(\delta z + Q_z\alpha) + \phi_2(\alpha, \delta))\},$$

where $\phi_2(\alpha, \delta) = \log\left(\int \exp\{Q_z\alpha + \delta z\}\right)$ and $\delta = \delta_1$. Typically, Y^{nmi} and Y^{mi} is positively

associated if they are associated, which restricts the sensitivity parameter δ to satisfying the inequality $\delta \geq 0$. For the relationship between $P(Y^{nmi} = 0|Y^{mi} = 1)$ and $P(Y^{nmi} = 0|Y^{mi} = 0)$, we use a sensitivity parameter τ as the ratio of the probabilities $\tau = P(Y^{nmi} = 0|Y^{mi} = 1)/P(Y^{nmi} = 0|Y^{mi} = 1)$. From the positive association of Y^{nmi} and Y^{mi}, the parameter τ should be above 1, i.e. $\tau \geq 1$. We conduct a sensitivity analysis by using the sensitivity parameter δ and τ in Section 5.5 to investigate the impact of the violation of Assumption 3 (ii) on the estimate of the MAFF.

5.3.4. Incorporating covariates

Another possible violation scenario of the independence assumption is that both Y^{nmi} and Y^{mi} are dependent on some variables. To make the assumption more plausible, we need to control for those confounding variables.

Let $\mathbf{X}$ be the $s \times 1$ vector. To incorporate covariates, consider the following model:

$$\begin{aligned} g_1(z|\mathbf{X} = \mathbf{x}; q, \alpha, \eta) &:= f(D^{no.nmi} = z|Y^{mi} = 0, \mathbf{X} = \mathbf{x}) \\ &= \frac{\exp(\mathbf{x}^T q)}{1 + \exp(\mathbf{x}^T q)} \cdot I(z = 0) \\ &+ \frac{1}{1 + \exp(\mathbf{x}^T q)} \cdot \exp\{(\mathbf{x}^T \eta) \cdot z + Q_z^T \alpha - \phi_1(\alpha, \eta)\} \cdot I(z > 0), \end{aligned}$$

where q and η are $s \times 1$ vectors and $\phi_1(\alpha, \eta) = \log\left\{\int \exp((\mathbf{x}^T \eta) \cdot z + Q_z^T \alpha)\right\}$. The parameter η represents dependence of the density on $\mathbf{X}$. As in Section 5.3.1, we assume an exponential tilt model for both z and $\mathbf{X}$. Therefore, the conditional density $f(D^{no.nmi} = z|Y^{mi} = 1, \mathbf{X} = \mathbf{x})$ is given as

$$\begin{aligned} g_2(z|\mathbf{X} = \mathbf{x}; \alpha, \eta, \gamma) &:= f(D^{no.nmi} = z|Y^{mi} = 1, \mathbf{X} = \mathbf{x}) \\ &= \exp\{\gamma_0 + (\mathbf{x}^T \gamma_1) \cdot z\} \cdot \exp\{(\mathbf{x}^T \eta) \cdot z + Q_z^T \alpha - \phi_1(\alpha, \eta)\} \\ &= \exp\{(\mathbf{x}^T (\eta + \gamma)) \cdot z + Q_z^T \alpha - \phi_2(\alpha, \gamma, \eta)\} \end{aligned}$$

where $\gamma = \gamma_1$ is a $s \times 1$ vector and $\phi_2(\alpha, \gamma, \eta) = \log\left\{\int \exp((\mathbf{x}^T (\eta + \gamma)) \cdot z + Q_z^T \alpha)\right\}$.

Besides these models for $g_1(z|\mathbf{X} = \mathbf{x})$ and $g_2(z|\mathbf{X} = \mathbf{x})$, the probability that a fever is observed is depend on $\mathbf{X} = \mathbf{x}$ and the mixing proportion λ^* is also depend on $\mathbf{X} = x$. We denote $p := p(\mathbf{x})$ as the probability $\Pr(Y^{obs} = 1|\mathbf{X} = \mathbf{x})$. Similarly, we use notation $\lambda^* := \lambda^*(\mathbf{x})$. The joint density given covariates, $(Y^{obs}, D^{obs}|\mathbf{X})$, can be represented as the product of $f(D^{obs}|Y^{obs}, \mathbf{X}) \times f(Y^{obs}|\mathbf{X})$. Therefore, similar to the likelihood (5.3.2), the log-likelihood function can be decomposed into two sub-loglikelihood functions,

$$\ell(p, \lambda^*, q, \alpha, \eta, \gamma) = \ell_1(p) + \ell_2(\lambda^*, \alpha, \eta, \gamma),$$

where

$$\ell_1(p) = \sum_{i=1}^{n} \log\{\Pr(Y^{obs} = Y_i^{obs} | \mathbf{X} = \mathbf{x}_i)\}$$

$$\ell_2(\lambda^*, q, \alpha, \eta, \gamma) = \sum_{i=1}^{n_0} \log\left\{\int h(D_{0i}^{obs}; z) \cdot g_1(z|\mathbf{x}_i; q, \alpha, \eta) dz\right\}$$
$$+ \sum_{i=1}^{n_1} \log\left\{(1-\lambda^*) \cdot \int h(D_{1i}^{obs}; z) \cdot g_1^*(z|\mathbf{x}_i; q, \alpha, \eta) dz\right.$$
$$\left. + \lambda^* \cdot \int h(D_{1i}^{obs}; z) g_2(z|\mathbf{x}_i; \alpha, \eta, \gamma) dz\right\}.$$

The sub-loglikelihood function $\ell_1(p)$ is only needed for estimating $p(\mathbf{x})$. Any regression method can be used to estimate the conditional probability. For example, nonparametric logistic regression methods can be used. However, the parameter $\lambda^*(\mathbf{x})$ is estimated simultaneously with other parameters from $\ell_2(\lambda^*, q, \alpha, \eta, \gamma)$. We restrict the model for this as $\lambda^*(\mathbf{x}) = \exp(\mathbf{x}^T\kappa)/(1+\exp(\mathbf{x}^T\kappa))$ where κ is a $s \times 1$ vector. Although incorporating covariates require more computation time, but the estimation scheme is the same as the estimation without covariates discussed in Section 5.3.2. We apply this approach to the Kilombero malaria data in Section 5.5.

5.4. Simulation Study

In this section, we evaluate the performance of our proposed method including the regular likelihood approach and the penalized likelihood approach in a simulation study. In addition to evaluating the performance of our proposed method, we compare it to the existing methods by considering various simulation settings. The distribution of g_1 is a mixture of a point mass at zero and a distribution for positive parasite levels and the distribution of g_2 satisfies $g_2(0) = 0$. We consider two scenarios; (i) g_1 and g_2 are exponential family distributions and (ii) g_1 and g_2 are not exponential family distributions. For the first scenario, we assume

$$g_1(z) = q \cdot I(z = 0) + (1-q) \cdot \mathrm{TN}_{(0,\infty)}(\mu_1, \sigma_1) \cdot I(z > 0)$$

where q is the proportion of zero parasite level and $\mathrm{TN}_{(0,\infty)}(\mu, \sigma)$ is a truncated normal distribution with mean μ and standard deviation σ in the interval $(0, \infty)$. The distribution of g_2 can only take positive parasite levels,

$$g_2(z) = \mathrm{TN}_{(0,\infty)}(\mu_2, \sigma_2).$$

The second scenario considers uniform distributions, which are not in exponential family distributions. To be specific, $g_1(z) = q_1 \cdot I(z = 0) + \big\{(1-q_1)q_2 \cdot \mathrm{TN}_{(0,\infty)}(\mu_1, \sigma_1) + (1-q_1)(1-q_2) \cdot U(0, 2\mu_1)\big\} \cdot I(z > 0)$ and $g_2(z) = q_2 \cdot \mathrm{TN}_{(0,\infty)}(\mu_2, \sigma_2) + (1-q_2) \cdot U(0, 2\mu_2)$ where $U(a, b)$ is the uniform distribution in the interval (a, b). Throughout the simulation study, the probability q_2 is fixed as $1/8$. The number of parameters is chosen as $dim(\alpha) = 3$ for g_1 in model (5.3.1). Along with the parameter q, the total number of parameters in $g_1(z)$ is 4. Similarly, including a exponential tilt parameter γ, the total number of parameter

Table 15: Exponential family distribution case. Means (standard deviations) of the estimators in simulation settings are displayed; P represents the power model regression method, S represents the adjusted semiparametric method, and LI represents the nonparametric method. True MAFF is 0.5

			MAFF				
n	q	β	Regular	Penalized	P	S	LI
500	0.2	1	0.501 (0.120)	0.488 (0.118)	0.443 (0.089)	0.471 (0.083)	0.475 (0.078)
		0.8	0.499 (0.117)	0.487 (0.112)	0.406 (0.102)	0.437 (0.089)	0.441 (0.081)
		0.2	0.512 (0.056)	0.491 (0.045)	0.119 (0.037)	0.138 (0.064)	0.143 (0.059)
	0.8	1	0.499 (0.028)	0.497 (0.028)	0.485 (0.027)	0.487 (0.026)	0.483 (0.025)
		0.8	0.498 (0.028)	0.495 (0.028)	0.483 (0.027)	0.485 (0.027)	0.480 (0.025)
		0.2	0.499 (0.024)	0.495 (0.022)	0.453 (0.025)	0.457 (0.024)	0.444 (0.020)
1000	0.2	1	0.499 (0.081)	0.490 (0.082)	0.440 (0.063)	0.468 (0.058)	0.472 (0.055)
		0.8	0.500 (0.076)	0.490 (0.072)	0.406 (0.072)	0.437 (0.062)	0.438 (0.057)
		0.2	0.507 (0.039)	0.495 (0.031)	0.117 (0.022)	0.138 (0.039)	0.135 (0.042)
	0.8	1	0.498 (0.020)	0.496 (0.020)	0.484 (0.019)	0.486 (0.019)	0.482 (0.018)
		0.8	0.498 (0.020)	0.497 (0.020)	0.483 (0.019)	0.485 (0.018)	0.480 (0.017)
		0.2	0.499 (0.017)	0.497 (0.016)	0.455 (0.018)	0.458 (0.017)	0.446 (0.015)

in $g_2(z)$ is 4. Also, we assume that we know the Poisson measurement error mechanism $D^{obs} \sim Pois(D^{cur})$, i.e. the mechanism h is the standard Poisson distribution.

In addition, we consider the following three factors that may affect the performance of our method.

1. Size of fever killing effect. Three different sizes of the fever killing effect are considered. The settings are large fever killing effect (fever kills approximately 80% of parasites which means $\beta = 0.2$), small fever killing effect (fever kills approximately 20% of parasites, $\beta = 0.8$) and no fever killing effect ($\beta = 1$). The no fever effect case will be used as a standard for comparison between the other two fever killing effects settings.

2. Endemicity. Endemic regions differ greatly by how many people have the malaria parasites in their blood. Endemicity could affect the variance of the estimate of the MAFF. Two levels of endemicity are considered: mesoendemic $q = 0.8$ (low to moderate) and holoendemic $q = 0.2$ (high).

3. Sample size n. Two sample sizes, $n = 500$ and $n = 1000$, are considered.

There are $3 \times 2 \times 2 = 12$ settings for each scenario to investigate the effect of the settings on the performance of our proposed method.

We use both the regular likelihood approach and the penalized likelihood approach to estimate the MAFF and compare them to the existing methods that do not account for fever killing and measurement error. Table 15 shows the means and standard deviations of the maximum likelihood estimates in the various settings when the true models are in the expo-

Table 16: Non-exponential family distribution case. Means (standard deviations) of the estimators in simulation settings are displayed; P represents the power model regression method, S represents the adjusted semiparametric method, and LI represents the nonparametric method. True MAFF is 0.5

			MAFF				
n	q	β	Regular	Penalized	P	S	LI
500	0.2	1	0.492 (0.116)	0.499 (0.126)	0.423 (0.094)	0.458 (0.087)	0.466 (0.081)
		0.8	0.488 (0.112)	0.503 (0.121)	0.379 (0.103)	0.420 (0.091)	0.430 (0.080)
		0.2	0.516 (0.062)	0.517 (0.071)	0.123 (0.033)	0.133 (0.060)	0.140 (0.055)
	0.8	1	0.498 (0.029)	0.501 (0.029)	0.480 (0.027)	0.482 (0.027)	0.478 (0.026)
		0.8	0.497 (0.030)	0.499 (0.031)	0.479 (0.027)	0.481 (0.026)	0.475 (0.025)
		0.2	0.501 (0.025)	0.505 (0.027)	0.451 (0.025)	0.454 (0.024)	0.441 (0.021)
1000	0.2	1	0.502 (0.091)	0.499 (0.095)	0.420 (0.065)	0.454 (0.060)	0.462 (0.055)
		0.8	0.498 (0.090)	0.498 (0.093)	0.379 (0.074)	0.421 (0.063)	0.430 (0.056)
		0.2	0.503 (0.051)	0.493 (0.040)	0.125 (0.022)	0.138 (0.038)	0.137 (0.040)
	0.8	1	0.497 (0.020)	0.497 (0.020)	0.481 (0.019)	0.483 (0.018)	0.478 (0.018)
		0.8	0.498 (0.022)	0.498 (0.022)	0.480 (0.019)	0.482 (0.019)	0.476 (0.018)
		0.2	0.501 (0.019)	0.498 (0.020)	0.450 (0.017)	0.453 (0.017)	0.440 (0.015)

nential family. The means and the standard deviations of the estimates are obtained from 1000 repetitions. Three aspects are found in this table. First, both the regular likelihood and the penalized likelihood approaches provide approximately unbiased estimates of the true MAFF while other existing methods (P, S, LI) are biased. There is a trend that the regular likelihood estimates have lower bias and slightly larger standard deviations than the penalized likelihood estimates. This is because the true models are exponential family distributions. Second, the larger proportion of zero parasite level contributes to more efficient estimates. That is, the estimate of the MAFF in mesoendemic regions is more efficient than the estimate in holoendemic regions. Finally, as sample size n increases, both the approaches produce estimates which are closer to the true MAFF value 0.5 (less bias) and have smaller standard deviations.

Table 16 displays the means and standard deviations of the estimates when the true models are not exponential family distributions. As can be seen in Table 16, a larger n and a larger q contribute to a smaller standard deviation. A smaller β leads to a smaller standard deviation, but it also leads to a bias. A different aspect in Table 16 compared to Table 15 is that the penalized likelihood approach generally performs better with smaller standard deviations than the regular likelihood approach in the non-exponential family distribution case of Table 16.

Furthermore, we examine the impact of misspecification of measurement error mechanism on estimation. Consider three measurement error mechanisms: (1) Poisson error (i.e., $D^{obs} \sim Pois(D^{cur})$), (2) Negative binomial error with $r = 10$ with mean D^{cur} (i.e., $D^{obs} \sim NB(D^{cur}, 10)$) and (3) Negative binomial error with $r = 20$ (i.e., $D^{obs} \sim NB(D^{cur}, 20)$). We note that if $r \to \infty$, the negative binomial distribution $NB(D^{cur}, r)$ converges to

Table 17: Simulations for misspecification of the measurement error model. True MAFF is 0.5.

		MAFF	
True	Used	Regular	Penalized
$Pois(D^{cur})$	$Pois(D^{cur})$	0.498 (0.078)	0.501 (0.077)
	$NB(10, D^{cur})$	0.444 (0.075)	0.478 (0.061)
	$NB(20, D^{cur})$	0.470 (0.085)	0.483 (0.080)
$NB(10, D^{cur})$	$Pois(D^{cur})$	0.537 (0.087)	0.535 (0.087)
	$NB(10, D^{cur})$	0.498 (0.080)	0.502 (0.078)
	$NB(20, D^{cur})$	0.518 (0.087)	0.518 (0.086)
$NB(20, D^{cur})$	$Pois(D^{cur})$	0.519 (0.078)	0.519 (0.077)
	$NB(10, D^{cur})$	0.477 (0.081)	0.490 (0.076)
	$NB(20, D^{cur})$	0.499 (0.077)	0.500 (0.076)

$Pois(D^{cur})$. Consider the setting $n = 1000, \beta = 1, q = 0.2$ used in the previous simulations. Table 17 shows simulation results for misspecification. In every simulation, the penalized likelihood method provides estimates closer to the true MAFF 0.5 than the regular likelihood method does.

We will use the penalized likelihood method for the analysis of the Kilombero malaria data in the next section.

5.5. Application to the Data From Kilombero, Tanzania

We consider data from a study of children in the Kilombero District (Morogoro Region) of Tanzania (Smith et al., 1994). The study collected parasite density levels and the presence of fever among 426 children under six years of age in two villages from June 1989 until May 1991 in the Kilombero District (Morogoro Region) of Tanzania. This area is highly endemic for *Plasmodium falciparum* malaria. A total of 1996 blood films from the 426 children were examined. Smith et al. (1994) found that the correlation between consecutive observations on the same child is not significant and the impact of the correlation on the MAFF is negligible. We will follow Smith et al. (1994) in assuming that the 1996 collected observations are independent and for brevity will describe the data as involving 1996 children. In this dataset, there are $n_1 = 137$ children who had a fever and $n_0 = 1859$ children who did not have a fever. The former is a group of febrile children whose fever was caused by either malaria infection or non-malaria infection. The latter is a group of afebrile controls that is used to provide information on the parasite density of the non-malaria infection population. Table 18 summarizes the data. The proportions of zero parasite level is $0.086 = 160/1858$ and $0.117 = 16/137$ in the afebrile and febrile groups respectively. The proportion of zero parasite density level in the afebrile group must be greater than that in the febrile group so the proportions in Table 18 implies some errors in measuring malaria parasites.

Also, in the absence of measurement error and fever killing, under Assumption 3, the

Table 18: Summary of the data from Kilombero, Tanzania

Parasite level	Afebrile	Febrile
$= 0$	160	16
> 0	1698	121
Total	1858	137

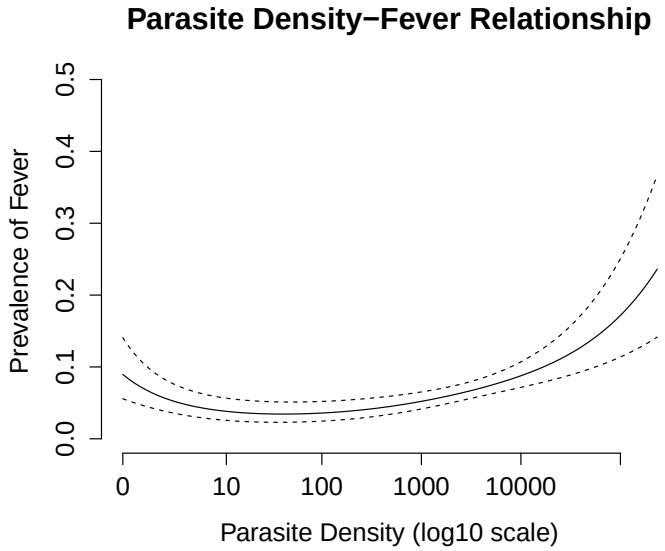

Figure 7: The relationship between parasite density and probability of fever. The solid curve represents the point estimate across parasite density obtained by using penalized splines, and the dashed curves are 95% pointwise confidence intervals.

probability of a fever should be monotonically increasing in the parasite density. Figure 7 shows the relationship between probability of fever and parasite density and also suggests a violation of the assumptions. Smith et al. (1994) point out that this phenomenon that the probability of a fever is not monotonically increasing in the parasite density has been observed in many other datasets, and consider it a consequence of non-malarial fevers suppressing low density parasitaemia, i.e., fever killing.

The existing four estimation methods are compared: (1) logistic regression method (L), (2) logistic regression with power (P), (3) local linear smoothing followed by isotonic regression (LI) and the semiparametric method (S). These methods do not account for the fever killing effect and measurement error problems. Under the assumption of the absence of the fever killing effect and measurement error, the estimates of the MAFF from the existing methods are 0.176 (L), 0.202 (P), 0.177 (LI) and 0.177 (S) shown in the bottom of Table 19. The displayed standard deviations are computed from 2000 bootstrapped samples.

For choosing the dimension of α, we conducted 10-fold likelihood-based cross-validation, and found that cross-validated values start to be flattened at $\dim(\alpha) = 4$. For the rest of

analysis, we use a model with a 4-dimensional parameter α. Also, to apply our proposed method, a measurement error mechanism needs to be specified. In the Kilombero study, the number of parasites is counted in a predetermined number of white blood cells (WBCs), usually 200, and then the parasite density per μl is estimated as 40 times the count, under the assumption that there are 8000 WBCs per μl. The simplest choice is the Poisson measurement error mechanism. The Poisson measurement error mechanism will hold if the only source of measurement error is the sampling of parasites from 1/40 μl of blood and parasites are uniformly distributed throughout the blood. Without any other sources of measurement error but this sampling error, the measurement error model would be (M1) $h_1(D^{obs}|D^{cur}) \sim Poisson(D^{obs}/40; D^{cur}/40)$. As we discussed in Section 5.2.2, microscopy error is another source of measurement error as well as sampling error. A more complicated model than the Poisson model to account for the microscopy error is a negative binomial measurement error model. The negative binomial (NB) model is (M2) $h_2(D^{obs}|D^{cur}) \sim \text{NB}(D^{obs}/40; r, r/(r + D^{cur}/40))$ where r is the dispersion parameter. From O'Meara et al. (2007), we estimate the dispersion parameter is $r = 6$. The estimation process is shown in Appendix A.4.2. Furthermore, another source of measurement error is WBCs count variability; the number of WBCs per μl is not fixed as 8000, but varies from person to person (McKenzie et al., 2005). Based on McKenzie et al. (2005), we consider a discrete distribution $w(z)$ of WBCs counts per μl: Point masses of $(4, 5, 6, 7, 8, 9, 10, 11, 12) \times 10^3 = (.12, .16, .20, .16, .16, .10, .04, .04, .02)$. The distribution $w(z)$ accounts for the potential effect of the variability of WBCs counts. The most complicated measurement error model (M3) we consider combines the microscopy error and the WBC count variability, and is $h_3(D^{obs}|D^{cur}) \sim \sum w(z) \cdot \text{NB}(D^{obs}/(z/200); r, r/(r + D^{cur}/(z/200)))$. We note that the distribution h_1 has the smallest variance and the distribution h_3 has the largest variance. The models (M1)-(M3) are considered and compared in our analysis.

In addition to the measurement error mechanism, the size of fever killing $1 - \beta$ needs to be specified, but it is not known precisely based on current scientific knowledge. We consider a series of various plausible fever killing sizes, and calculate the corresponding estimates of the MAFF. Based on previous studies, we can shorten the plausible range of the size of fever killing effect by using children's temperature data. We found that the temperature data are distributed between 37.5°C and 40°C with 90% percentile 38.7°C and 95% percentile 39.1°C. Long et al. (2001) found that when the temperature is 38.8°C, the fever killing effect was 50% so we consider 50% an upper bound on the fever killing effect (i.e., the range of the fever killing size is $1 - \beta \in [0, 0.5]$). We do not consider no fever killing to be plausible but include it for comparison purposes. Furthermore, the assumption that there is a fixed size of fever killing across population can be eased by incorporating the temperature data into the analysis by accounting for temperature-varying fever killing size. However, incorporating the temperature data is beyond the scope of our paper.

Table 19 shows the estimates of the MAFF from the different values of the fever killing effect parameter β for each measurement error model from (M1) to (M3). As the fever killing effect becomes larger (i.e., β decreases), the estimate of the MAFF increases roughly from 0.18 to 0.34 for the considered measurement error mechanisms. This shows that the problem of fever killing effects is much severer in estimating the MAFF than the problem of measurement error is. This is because three mechanisms recovered the true parasite

Table 19: Estimates of the MAFF. The upper table: the estimates corresponding to the different sizes of fever killing; $1-\beta = 0.5$ means 50% fever killing and $1-\beta = 0$ means no fever killing. The standard deviations are computed from 1000 bootstrapped samples. The lower table: the estimates from the existing methods.

	MAFF		
$1-\beta$	M1	M2	M3
0.00	0.187 (0.063)	0.190 (0.059)	0.189 (0.057)
0.05	0.201 (0.059)	0.200 (0.061)	0.203 (0.060)
0.10	0.215 (0.063)	0.218 (0.058)	0.218 (0.059)
0.15	0.231 (0.065)	0.234 (0.062)	0.233 (0.059)
0.20	0.247 (0.067)	0.251 (0.062)	0.249 (0.063)
0.25	0.264 (0.067)	0.268 (0.060)	0.266 (0.064)
0.30	0.282 (0.071)	0.286 (0.062)	0.284 (0.065)
0.35	0.302 (0.068)	0.306 (0.067)	0.303 (0.065)
0.40	0.323 (0.069)	0.327 (0.065)	0.324 (0.069)
0.45	0.344 (0.074)	0.349 (0.069)	0.345 (0.066)
0.50	0.368 (0.072)	0.373 (0.071)	0.369 (0.072)

	L	P	S	LI
MAFF	0.176 (0.042)	0.202 (0.074)	0.177 (0.063)	0.182 (0.079)

density similarly, especially, the proportion of zero parasite. Even in the positive number of parasites, no parasite can be observed due to measurement errors. We found that the recovered proportion of zero parasite was almost identical in every considered measurement error mechanism, which leads to similar MAFF estimates.

In the analysis of the Tanzania malaria data, an additional age variable can be used to make Assumption 3 more plausible. It is studied that there is evidence that younger children tend to have a higher probability to have a malarial fever (Rogier et al., 1996) as well as a non-malarial fever. We control for this variable using the method discussed in Section 5.3.4. Figure 8 visually shows the estimates after incorporating the age variable compared to the results in Table 19. The figure plots the estimates of the MAFF on the size of fever killing effects for the mechanisms (M1) to (M3). One distinct pattern is that the MAFF curves are slightly tilted after incorporating the age variable, and the difference in the MAFF estimates between before and after incorporating the age variable is at most 0.017. Also, the three measurement error mechanisms produced similar estimates of the MAFF for each setting. This analysis shows that the violation of Assumption 3 due to age is not significant.

We also conduct an additional sensitivity analysis for violation of Assumption 3. We use two sensitivity parameters δ_1 and τ discussed in Section 5.3.3. The sensitivity parameter δ_1 represents how two densities $f(D^{no.nmi} = z|Y^{mi} = 0, Y^{nmi} = 0)$ and $f(D^{no.nmi} = z|Y^{mi} = 0, Y^{nmi} = 1)$ differ ($\delta_1 = 0$ means that the densities are equal), and the sensitivity parameter τ represents the relative rate of not having a non-malarial fever according to Y^{mi}. We consider plausible ranges of the parameters: $0 \leq \delta_1 \leq 0.1$ and $1 \leq \tau \leq 1.06$. Figure 9

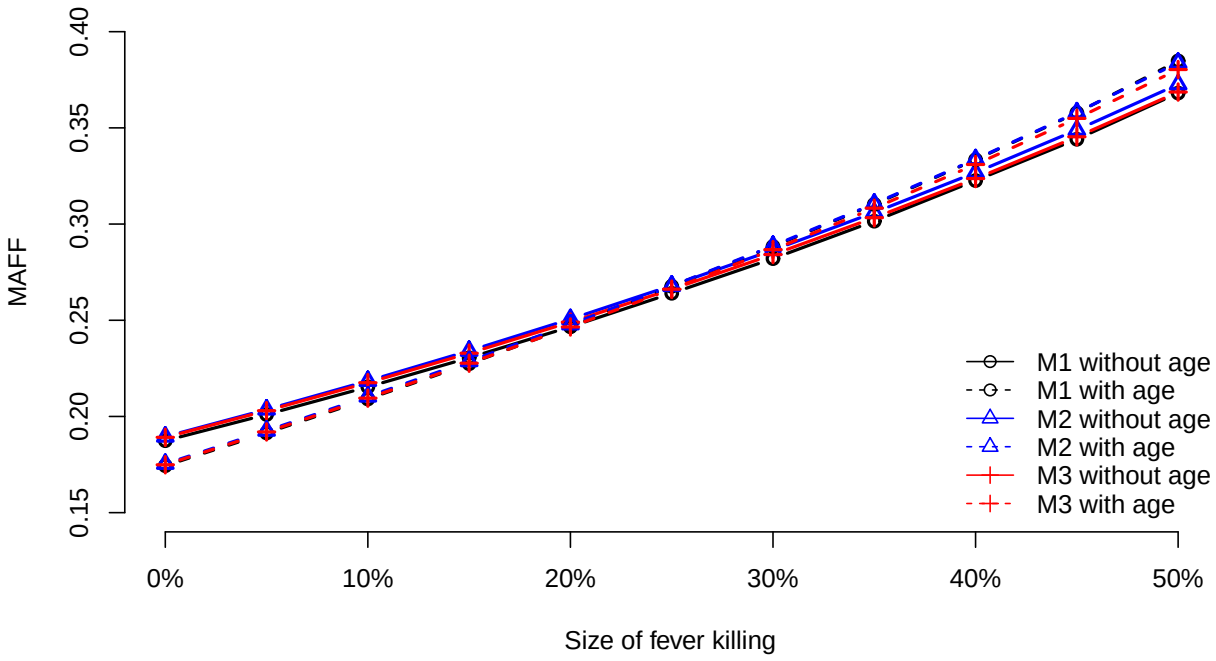

Figure 8: The plot of the estimates of the MAFF on the size of fever killing (i.e., $100(1-\beta)\%$) for the measurement error models (M1)-(M3).

shows the estimates of the MAFF according to the values of δ_1 and τ for a size of fever killing β ($\beta = 0, 0.2$ or 0.5) and a measurement error mechanism M1 (M2 and M3 provide similar results). The estimates are plotted by using contour plots. The effect of deviation from Assumption 3 (ii) is shown differently according to which factor caused a violation. As δ_1 increases, the estimate of the MAFF decreases and as τ increases, the estimate increases. Therefore, a mixed effect of δ_1 and τ appears to cancel each other out and to have a slight impact on the estimate. Another noticeable pattern is that the impact of δ_1 is more severe than the impact of τ when a fever killing effect is small, but this pattern is reversed when a fever killing effect is large.

5.6. Summary

In this article, we have proposed a new approach to estimate the MAFF in the presence of both fever killing and measurement error. We have shown that existing MAFF estimators can be substantially biased in the presence of these problems. We develop a new estimator using the g-modeling approach to the Bayes deconvolution problem. To develop this new estimator, we extended the existing g-modeling approach that solves the convolution problem in non-mixture data to a setting of two-component mixture data such as malaria data. Under the assumptions that the size of fever killing effect is known and measurement

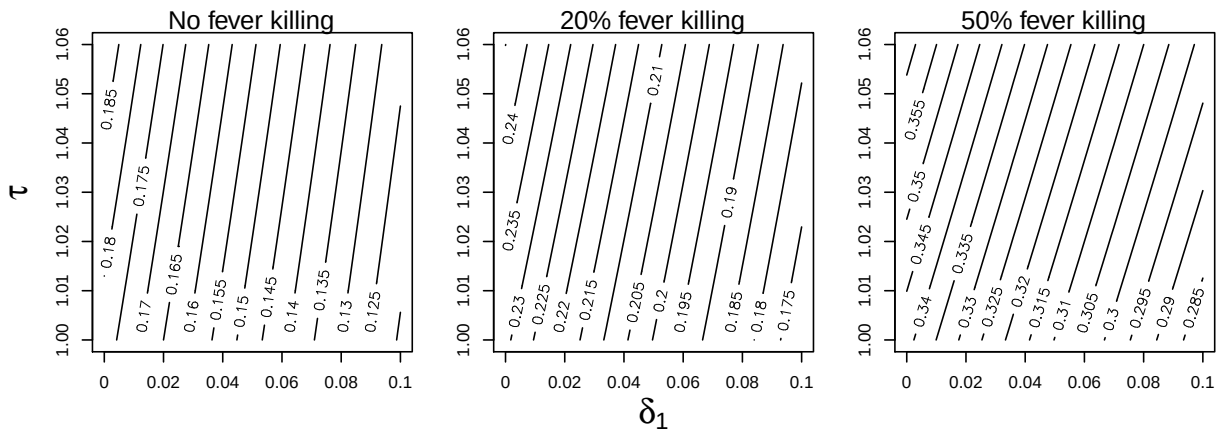

Figure 9: Sensitivity Analysis for violation of Assumption 3. From left to right, the size of fever killing is from 0% to 50%. The estimates of the MAFF are represented as contour levels according to the values of δ_1 and τ. The difference between adjacent contour levels is 0.005 in every sensitivity analysis.

error mechanism is known, our new estimator performs well. In practice, when the size of the fever killing effect is not known, we recommend choosing a plausible range of the fever killing size and comparing the corresponding estimates for a sensitivity analysis. To get a better estimate of the MAFF, further research is needed to better understand the fever killing effect to be able to get a narrower range of the fever killing size. Another difficulty in practice is to specify the measurement error mechanism. This problem can be eased by considering several plausible measurement error mechanism models from the simplest to the most complicated (i.e., allowing more sources of measurement errors) as we did in Section 5.5. If a more complicated model produces a similar estimate to that of a less complicated model, then we can be more confident with our conclusion. More research on understanding the measurement error mechanism would also be useful.

CHAPTER 6 : Discussions

Throughout Chapter 2 to 5, we studied three other quantities of interest which are more complicated than the ATE. In Chapter 2, we studied the distributional treatment effects by estimating potential outcome distributions. We showed the limitations of the existing estimation methods, and provided a new nonparametric likelihood method to overcome these limitations. We further developed a theory on the nonparametric BL to test whether a treatment caused any change in the potential outcome distributions. In Chapter 3 and 4, we developed novel approaches of discovering effect modification in a data-driven way while avoiding the issue of using the data twice. More specifically, the CART approach was applied to the absolute treated-minus-control pair differences that do not use treatment assignment. The CART method can deal with high-dimensional observational study data, and can efficiently detect effect modification despite the lack of supporting statistical properties. Furthermore, we developed a statistical approach based on multiple testing correction with providing statistical properties. In Chapter 6, we established a theory of the MAFF in causal inference. We elucidated the causal definition of the MAFF using the potential outcome framework and developed a method to estimate the MAFF from the data. We delineated potential measurement error scenarios and the problems that arise in the presence of such measurement errors. To obtain an unbiased estimate of the MAFF, we developed a novel maximum likelihood estimation method to incorporate the potential measurement errors based on exponential family g-modeling.

In conclusion, the focus of the current theory and methods in observational studies has been restricted to the ATE. We studied other treatment effects beyond the ATE, and revealed that more complex and sophisticated causal inference can be made with the same underlying assumptions.

APPENDIX

A.1. Simulation for the Estimation of Distributions from Chapter 2

Table 20: Normal Mixture. The average performance comparison between the MBL method, the method of moment method and the parametric normal mixture method when the true distributions are normal; AD means the average discrepancy from the true CDF.

			MBL		MOM		Parametric	
Causal effect	IV	Z	AD	SE	AD	SE	AD	SE
No	Strong	$Z=0$	0.0030	0.0028	0.0031	0.0029	0.0016	0.0023
No	Strong	$Z=1$	0.0030	0.0028	0.0031	0.0029	0.0016	0.0023
Some	Strong	$Z=0$	0.0030	0.0030	0.0031	0.0031	0.0017	0.0026
Some	Strong	$Z=1$	0.0030	0.0028	0.0031	0.0029	0.0017	0.0025
No	Weak	$Z=0$	0.0287	0.0307	0.0951	0.7838	0.0182	0.0287
No	Weak	$Z=1$	0.0274	0.0312	0.0934	0.8526	0.0188	0.0312
Some	Weak	$Z=0$	0.0277	0.0288	0.0764	0.5993	0.0176	0.0269
Some	Weak	$Z=1$	0.0301	0.0343	0.0773	0.3546	0.0184	0.0272

We conduct a simulation study to examine the accuracy of our proposed MBL estimation method. We consider three different methods to estimate the outcome CDFs for compliers; the MBL method, the MOM method described in Section 2.2.2 and a parametric normal mixture method Imbens and Rubin (1997). The parametric normal mixture model assumes that all outcome distributions for compliance classes have normal distributions. Then, using the EM algorithm, it estimates the means and the variances of the outcome distributions. Specifically, we consider two simulation scenarios: (1) normal mixture models and (2) gamma mixture models.

Also, in each scenario, we consider two more factors that can affect the performance of these three methods. First, we consider whether there is any effect of the treatment for compliers, i.e. whether the outcome distributions of $F_{co}^{(0)}$ and $F_{co}^{(1)}$ are the same or not the same. Second, we consider whether the IV is strong or weak. The strength of an IV is how strongly the IV is associated with the treatment. One common definition of a weak IV is that the first stage F-statistic when the treatment is regressed on the IV is less than 10 Stock et al. (2002). We consider a strong IV setting where the proportions of subpopulations (co, nt, at) is (1/3, 1/3, 1/3) (average first stage F-statistic$\approx$124) and a weak IV setting where proportions of $(co, nt, at) = (0.10, 0.45, 0.45)$ (average first stage F-statistic$\approx$10).

We repeat simulations for 1000 times with the sample size $n = 1000$ to see average performance of estimating the true CDFs. The measurement of the discrepancy between the estimated and the true CDFs is defined by L_2 distance, i.e. if the true CDF is F and our

estimated CDF is $\hat{F}$, the L_2 distance is

$$L_2(F, \hat{F}) = \int \{F(x) - \hat{F}(x)\}^2 dF(x). \tag{A.1.1}$$

Under the assumption that all subpopulations have normal outcome distributions, we consider the case that there is no causal effect of Z on Y for compliers, specifically, $F_{co}^{(0)} = F_{co}^{(1)} \sim N(0, 4^2), F_{nt} \sim N(2, 4^2), F_{at} \sim N(-2, 4^2)$ and the case that there is some effect of Z on Y, specifically, $F_{co}^{(0)} \sim N(1, 4^2), F_{co}^{(1)} \sim N(-1, 4^2), F_{nt} \sim N(2, 4^2), F_{at} \sim N(-2, 4^2)$. Table 20 shows that the average performance of the MBL method is better than that of the MOM method in all cases, and it is not much different from the average performance of the parametric normal mixture method that assumes the correct parametric distributions. All three method are greatly affected by the IV being weak. In particular, the MOM method is more sensitive to the IV being weak; the MOM method and the MBL method perform similarly when the IV is strong but the MBL method is much better than the MOM method when the IV is weak.

Table 21: Gamma Mixture case. The average performance comparison between the MBL method, the MOM method and the parametric normal mixture method when the true distributions are nonnormal; AD means the average discrepancy from the true CDF.

			MBL		MOM		Parametric	
Causal effect	IV	Z	AD	SE	AD	SE	AD	SE
No	Strong	$Z = 0$	0.0030	0.0029	0.0031	0.0030	0.0054	0.0128
No	Strong	$Z = 1$	0.0029	0.0029	0.0030	0.0030	0.0061	0.0172
Some	Strong	$Z = 0$	0.0029	0.0027	0.0030	0.0028	0.0148	0.0443
Some	Strong	$Z = 1$	0.0028	0.0027	0.0029	0.0027	0.0198	0.0608
No	Weak	$Z = 0$	0.0290	0.0310	0.1048	0.9679	0.0888	0.1239
No	Weak	$Z = 1$	0.0271	0.0312	0.1116	1.6740	0.0753	0.1266
Some	Weak	$Z = 0$	0.0280	0.0294	0.0523	0.1292	0.0711	0.1173
Some	Weak	$Z = 1$	0.0271	0.0285	0.0508	0.1039	0.1047	0.1371

Under the normal assumption, the normal parametric mixture method is the best method of the three methods. However, the assumption of normality is a strong assumption. If normality does not hold, then the normal parametric method is no longer the best. Table 21 summarizes the simulation results of a gamma case. Let $\Gamma(\alpha, \beta)$ be a Gamma distribution with shape α and rate β. In Table 21, similar to the normal case, we consider the no causal effect case that $F_{co}^{(0)} = F_{co}^{(1)} \sim \Gamma(1.2, 1), F_{nt} \sim N(1, 1), F_{at} \sim N(1.4, 1)$ and the some causal effect case that $F_{co}^{(0)} \sim \Gamma(1.1, 1), F_{co}^{(1)} \sim \Gamma(1.3, 1), F_{nt} \sim N(1, 1), F_{at} \sim N(1.4, 1)$.

Table 21 shows that the MBL method is the dominant method in all scenarios considered when normality is not satisfied. Though the MOM method has a similar performance in

the strong IV setting, it is much worse than the MBL method in the weak IV setting as in the normal mixture model setting. Also, the parametric normal mixture method has significantly increased ADs with large SEs. In short, since the MBL method does not rely on any assumption about the distribution of the data, it is robust to any distribution assumption and is the least sensitive to the IV being weak.

A.2. Proofs from Chapter 2

A.2.1. Proof of Theorem 2.3.1

Observation A.2.1. *Fix $a \in [0,1]$. Then for every $x \in [0,1]$, $J(a,x) = a\log x + (1-a)\log(1-x) \leq a\log a + (1-a)\log(1-a) = J(a,a)$.*

Proof. The inequality is trivially satisfied when $a \in \{0,1\}$. Therefore, assume that $a \in (0,1)$, and define a random variable W which takes values $\frac{x}{a}$ and $\frac{1-x}{1-a}$ with probabilities a and $1-a$, respectively. Note that $\mathbb{E}W = 1$. Then by Jensen's inequality, $\mathbb{E}(\log W) = a\log\frac{x}{a} + (1-a)\log\frac{1-x}{1-a} \leq \log \mathbb{E}W = 0$, which completes the proof of the result. □

Lemma A.2.1. *Let $\boldsymbol{F} = (F_{co}^{(0)}, F_{nt}, F_{co}^{(1)}, F_{at})$ be as defined in (2.2.2). Then,*

$$\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta}_+} \mathbb{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta}} \mathbb{M}(\boldsymbol{\theta}) = \boldsymbol{F}, \tag{A.2.1}$$

Proof. By Observation A.2.1,

$$\mathbb{M}(\boldsymbol{\theta}) \leq \frac{1}{n}\sum_{b=1}^{n}\left\{\eta_{00}I(F_{00}(Y_b)) + \eta_{10}I(F_{10}(Y_b)) + \eta_{01}I(F_{01}(Y_b)) + \eta_{11}I(F_{11}(Y_b))\right\}. \tag{A.2.2}$$

Moreover, the equality is attained when $(\theta_{00},\theta_{10},\theta_{01},\theta_{11}) = (F_{00},F_{10},F_{01},F_{11})$. Recall that $F_{01}(t) = F_{at}(t)$, $F_{10}(t) = F_{nt}(t)$. Moreover, from (2.2.4) and (2.2.5), $F_{00}(t) = \frac{\phi_c}{\phi_c+\phi_n}F_{co}^{(0)}(t) + \frac{\phi_n}{\phi_c+\phi_n}F_{nt}(t)$ and $F_{11}(t) = \frac{\phi_c}{\phi_c+\phi_a}F_{co}^{(1)}(t) + \frac{\phi_a}{\phi_c+\phi_a}F_{at}(t)$. This implies that the equality in (A.2.2) above, is attained when $(\theta_{co}^{(0)},\theta_{nt},\theta_{co}^{(1)},\theta_{at}) = (F_{co}^{(0)},F_{nt},F_{co}^{(1)},F_{at})$. Finally, since $(F_{co}^{(0)},F_{nt},F_{co}^{(1)},F_{at}) \in \vartheta_+$, (A.2.1) follows. □

Observation A.2.2. *Fix $0 < t < 1$. Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. samples with distribution function $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$. Then, for $u,v \in \{0,1\}$*

$$F_{uv}(Y_{\lceil nt\rceil}) \xrightarrow{P} H_{uv}^{-1}(t),$$

where $H_{uv}(t) = \eta_{00}(F_{00}\circ F_{uv}^{-1})(t) + \eta_{01}(F_{01}\circ F_{uv}^{-1})(t) + \eta_{10}(F_{10}\circ F_{uv}^{-1})(t) + \eta_{11}(F_{11}\circ F_{uv}^{-1})(t)$.

Proof. Without of generality, take $u = 0$ and $v = 0$. Then the distribution of

$$W_1 := F_{00}(Y_1),\ W_2 := F_{00}(Y_2), \ldots,\ W_n := F_{00}(Y_n)$$

are i.i.d. samples with distribution function $H_{00}(t) = \eta_{00}t + \eta_{01}(F_{01} \circ F_{00}^{-1})(t) + \eta_{10}(F_{10} \circ F_{00}^{-1})(t) + \eta_{11}(F_{11} \circ F_{00}^{-1})(t)$. This implies, for any $0 < t < 1$, $F_{00}(Y_{(\lceil nt \rceil)}) = W_{(\lceil nt \rceil)} \xrightarrow{P} H_{00}^{-1}(t)$, where the last step uses the convergence of sample quantiles to the corresponding population quantiles. □

Lemma A.2.2. *Let $\mathbb{M}_n(\cdot)$ and $\mathbb{M}(\cdot)$ be as in (2.3.6) and (2.3.7). Then*

$$\sup_{\boldsymbol{\theta} \in \vartheta_+} |\mathbb{M}_n - \mathbb{M}|(\boldsymbol{\theta}) \xrightarrow{P} 0.$$

Proof. Denote $\theta_{01} = \theta_{at}$ and $\theta_{10} = \theta_{nt}$. Therefore, for $u, v \in \{0, 1\}$,

$$\begin{aligned} & T_{uv}^{(n)}(\theta_{uv}|\boldsymbol{Y}) - T_{uv}(\theta_{uv}|\boldsymbol{Y}) \\ &= \frac{1}{n} \cdot \sum_{b \in I_\kappa} \frac{n_{uv}}{n} \left\{ \mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}) \right\} \log \frac{\theta_{uv}(Y_{(b)})}{1 - \theta_{uv}(Y_{(b)})} \end{aligned}$$

and

$$\begin{aligned} & (\mathbb{M}_n - \mathbb{M})(\boldsymbol{\theta}) \\ &= \sum_{u,v \in \{0,1\}} \left(T_{uv}^{(n)}(\theta_{uv}|\boldsymbol{Y}) - T_{uv}(\theta_{uv}|\boldsymbol{Y}) \right) \\ &= \frac{1}{n} \sum_{u,v \in \{0,1\}} \sum_{b \in I_\kappa} \frac{n_{uv}}{n} \left\{ \mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}) \right\} \log \frac{\theta_{uv}(Y_{(b)})}{1 - \theta_{uv}(Y_{(b)})}. \end{aligned} \tag{A.2.3}$$

Now, by the monotonicity of θ_{uv},

$$\left| \log \frac{\theta_{uv}(Y_{(b)})}{1 - \theta_{uv}(Y_{(b)})} \right| \leq \left| \log \frac{\theta_{uv}(Y_{(\lceil n\kappa \rceil)})}{1 - \theta_{uv}(Y_{(\lceil n\kappa \rceil)})} \right| + \left| \log \frac{\theta_{uv}(Y_{(\lceil n(1-\kappa) \rceil)})}{1 - \theta_{uv}(Y_{(\lceil n(1-\kappa) \rceil)})} \right| = O_P(1),$$

and

$$\begin{aligned} & |F_{uv}(Y_{(b)}) \log \theta_{uv}(Y_{(b)}) + (1 - F_{uv}(Y_{(b)})) \log(1 - \theta_{uv}(Y_{(b)}))| \\ & \leq \max \left(|\log \theta_{uv}(Y_{\lceil n\kappa \rceil})|, |\log(1 - \theta_{uv}(Y_{\lceil n(1-\kappa) \rceil}))| \right) = O_p(1) \end{aligned}$$

Using this and $\sup_{t \in \mathbb{R}} |\mathbb{F}_{uv}(t) - F_{uv}(t)| = O_P(1/\sqrt{n})$, it follows from (A.2.3), that $|(\mathbb{M}_n - \mathbb{M})(\boldsymbol{\theta})| \leq O_P(1/\sqrt{n})$, completing the proof of the lemma. □

Lemma A.2.3. *Let $\boldsymbol{F} = \arg\max_{\boldsymbol{\theta} \in \vartheta_+} \mathbb{M}(\boldsymbol{\theta})$. Then*

$$\max_{\boldsymbol{\theta} \in B(\boldsymbol{F}, \delta)} \mathbb{M}(\boldsymbol{\theta}) < \mathbb{M}(\boldsymbol{F},),$$

where $B(\boldsymbol{F}, \delta) := \left\{ \boldsymbol{\theta} \in \vartheta_+ : \frac{1}{n} \sum_{b \in I_\kappa} ||\boldsymbol{\theta}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 > \delta_1 \right\}$.

Proof. Note that

$$\begin{aligned}
&\mathbb{M}(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{F}) \\
&= \sum_{u,v\in\{0,1\}} T_{uv}(\theta_{uv}|\boldsymbol{Y}) - T_{uv}(F_{uv}|\boldsymbol{Y}) \\
&= \frac{1}{n} \sum_{u,v\in\{0,1\}} \sum_{b\in I_\kappa} \eta_{uv} \left\{ F_{uv}(Y_{(b)}) \log \frac{\theta_{uv}(Y_{(b)})}{F_{uv}(Y_{(b)})} + (1 - F_{uv}(Y_{(b)})) \log \frac{1-\theta_{uv}(Y_{(b)})}{1-F_{uv}(Y_{(b)})} \right\}.
\end{aligned} \tag{A.2.4}$$

For a given $a \in (0,1)$, let $f_a(x) = a \log \frac{x}{a} + (1-a)\frac{1-x}{1-a}$. By a second order Taylor expansion around the point a, $f_a(x) = \frac{1}{2}(x-a)^2 f_a''(\gamma_{x,a})$ where $\gamma_{x,a} \in [x \wedge a, x \vee a]$ ($x \wedge a := \min(x,a)$ and $x \vee a := \max(x,a)$) and $f_a''(x) = -\frac{a}{x^2} - \frac{1-a}{(1-x)^2}$. Note that, for $a \in (0,1)$ fixed, the function $f_a''(x)$ is convex. It is easy to check that the minimum is attained at $x_0 = (\frac{a}{1-a})^{\frac{1}{3}}$, and $f_a''(x) \geq f_a''(x_0) > 1$. Then,

$$\begin{aligned}
\mathbb{M}(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{F}) &< -\frac{1}{n} \left\{ \sum_{u,v\in\{0,1\}} \sum_{b\in I_\kappa} \eta_{uv} \frac{(\theta_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}))^2}{2} \right\} \\
&\lesssim -\frac{1}{n} \sum_{b\in I_\kappa} \|\boldsymbol{\theta}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})\|_2^2 \\
&\leq -\delta,
\end{aligned}$$

completing the proof of the lemma. □

The proof of Theorem 2.3.1 can now be completed using the Lemma A.2.2 and A.2.3 above, as follows: By definition, $\mathbb{M}_n(\hat{\boldsymbol{\theta}}) \geq \sup_{\boldsymbol{\theta}\in\vartheta_+} \mathbb{M}_n(\boldsymbol{\theta})$.

By the definition of the BL estimates, $\mathbb{M}_n(\hat{\boldsymbol{\theta}}) \geq \mathbb{M}_n(\boldsymbol{F}) - o_P(1)$. By Lemma A.2.2, this implies that $\mathbb{M}_n(\hat{\boldsymbol{\theta}}) \geq \mathbb{M}(\boldsymbol{F}) - o_P(1)$. Therefore,

$$\begin{aligned}
\mathbb{M}(\boldsymbol{F}) - \mathbb{M}(\hat{\boldsymbol{\theta}}) &\leq \mathbb{M}_n(\hat{\boldsymbol{\theta}}) - \mathbb{M}(\hat{\boldsymbol{\theta}}) + o_P(1) \\
&\leq \sup_{\boldsymbol{\theta}\in\vartheta_+} |\mathbb{M}_n - \mathbb{M}|(\boldsymbol{\theta}) + o_P(1) \xrightarrow{P} 0.
\end{aligned} \tag{A.2.5}$$

By the Lemma A.2.3, for every $\delta > 0$ there exists $\varepsilon = \varepsilon(\delta) > 0$ such that $\mathbb{M}(\boldsymbol{\theta}) < \mathbb{M}(\boldsymbol{F}) - \varepsilon$ for every $\boldsymbol{\theta} \in B(\boldsymbol{F}, \delta)$ where $B(\boldsymbol{F}, \delta)$ is defined in Lemma A.2.3. Thus, the event $\{\boldsymbol{\theta} \in B(\boldsymbol{F}, \delta)\}$ is contained in the event $\{\mathbb{M}(\hat{\boldsymbol{\theta}}) < \mathbb{M}(\boldsymbol{F}) - \varepsilon\}$, and by (A.2.5),

$$\mathbb{P}(B(\boldsymbol{F}, \delta)) = \mathbb{P}\left(\frac{1}{n} \sum_{b\in I_\kappa} ||\hat{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 > \delta \right) \to 0,$$

This completes the proof of (2.3.8).

A.2.2. Proof of Theorem 2.3.2

Hereafter, denote $\theta_{01} = \theta_{at}$ and $\theta_{10} = \theta_{nt}$. Then from (A.2.3)

$$(\mathbb{M}_n - \mathbb{M})(\boldsymbol{\theta}) = \frac{1}{n} \sum_{u,v \in \{0,1\}} \frac{n_{uv}}{n} \sum_{b=1}^{n} \{\mathbb{F}_{uv}(Y_b) - F_{uv}(Y_b)\} \log \frac{\theta_{uv}(Y_b)}{1 - \theta_{uv}(Y_b)}.$$

Recall that $\tilde{\boldsymbol{\theta}} \in \vartheta_+$ is a function from $\mathbb{R} \to [0,1]^4$ such that $\tilde{\theta}(t) = (\theta_{co}^{(0)}(t), \theta_{nt}(t), \theta_{co}^{(1)}(t), \theta_{at}(t))'$. Moreover, recall that $F(t) = (F_{co}^{(0)}(t), F_{nt}(t), F_{co}^{(1)}(t), F_{at}(t))'$, are the true population distribution functions.

Define a matrix $\boldsymbol{V}_n$ as $4n \times 4n$ matrix such as $\boldsymbol{V}_n = \begin{pmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_n \end{pmatrix}$ where

$$\Sigma_b = \begin{pmatrix} \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(0)}(Y_b))^2} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(0)}(Y_b))(\partial \theta_{nt}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(0)}(Y_b))(\partial \theta_{co}^{(1)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(0)}(Y_b))(\partial \theta_{at}(Y_b))} \\ \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{nt}(Y_b))(\partial \theta_{co}^{(0)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{nt}(Y_b))^2} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{nt}(Y_b))(\partial \theta_{co}^{(1)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{nt}(Y_b))(\partial \theta_{at}(Y_b))} \\ \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(1)}(Y_b))(\partial \theta_{co}^{(0)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(1)}(Y_b))(\partial \theta_{nt}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(1)}(Y_b))^2} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{co}^{(1)}(Y_b))(\partial \theta_{at}(Y_b))} \\ \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{at}(Y_b))(\partial \theta_{co}^{(0)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{at}(Y_b))(\partial \theta_{nt}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{at}(Y_b))(\partial \theta_{co}^{(1)}(Y_b))} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\theta})}{(\partial \theta_{at}(Y_b))^2} \end{pmatrix} \Big|_{\boldsymbol{\theta}=\boldsymbol{F}} \tag{A.2.6}$$

From direct computation, the inverse matrix $\boldsymbol{V}_n^{-1} = \begin{pmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_n^{-1} \end{pmatrix}$ where

$$\Sigma_b^{-1} = n \begin{pmatrix} -(\frac{1-\lambda_0}{\lambda_0})^2 \frac{1}{Q_{10}(Y_b)} - \frac{1}{\lambda_0^2} \frac{1}{Q_{00}(Y_b)} & \frac{1-\lambda_0}{\lambda_0} \frac{1}{Q_{10}(Y_b)} & 0 & 0 \\ \frac{1-\lambda_0}{\lambda_0} \frac{1}{Q_{10}(Y_b)} & -\frac{1}{Q_{10}(Y_b)} & 0 & 0 \\ 0 & 0 & -(\frac{1-\lambda_1}{\lambda_1})^2 \frac{1}{Q_{01}(Y_b)} - \frac{1}{\lambda_1^2} \frac{1}{Q_{11}(Y_b)} & \frac{1-\lambda_1}{\lambda_1} \frac{1}{Q_{01}(Y_b)} \\ 0 & 0 & \frac{1-\lambda_1}{\lambda_1} \frac{1}{Q_{01}(Y_b)} & -\frac{1}{Q_{01}(Y_b)} \end{pmatrix} \tag{A.2.7}$$

where

$$
\begin{aligned}
Q_{00}(Y_b) &= \eta_{00}\frac{1}{F_{00}(Y_b)(1-F_{00}(Y_b))}\\
Q_{01}(Y_b) &= \eta_{01}\frac{1}{F_{01}(Y_b)(1-F_{01}(Y_b))}\\
Q_{10}(Y_b) &= \eta_{10}\frac{1}{F_{10}(Y_b)(1-F_{10}(Y_b))}\\
Q_{11}(Y_b) &= \eta_{11}\frac{1}{F_{11}(Y_b)(1-F_{11}(Y_b))}
\end{aligned}
\tag{A.2.8}
$$

The matrix $\boldsymbol{V}_n(Y_{(b)})$ is represented as $\Sigma_{(b)}$. The corresponding the inverse matrix $\boldsymbol{V}_n(Y_{(b)})^{-1}$ is given by $\Sigma_{(b)}^{-1}$.

Lemma A.2.4. *Let $\tilde{\boldsymbol{\theta}} \in \vartheta_+$ be such that $\frac{1}{n}\sum_{b\in I_\kappa}||\tilde{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 = O_P(1/n)$ and $||\tilde{\boldsymbol{\theta}}(Y_{(\lceil n\kappa\rceil)}) - \boldsymbol{F}(Y_{(\lceil n\kappa\rceil)})||_2 = o_P(1)$. Then*

$$
\begin{aligned}
(\mathbb{M}_n(\tilde{\boldsymbol{\theta}}) - \mathbb{M}_n(\boldsymbol{F})) &= \frac{1}{n^{\frac{3}{2}}}\sum_{b\in I}\langle\tilde{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)}), \boldsymbol{Z}_n(Y_{(b)})\rangle\\
&\quad + \frac{1}{2}\cdot\frac{1}{n}\sum_{b\in I_\kappa}(\tilde{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)}))'\boldsymbol{V}_n(Y_{(b)})(\tilde{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})) + O_P(n^{-\frac{3}{2}}),
\end{aligned}
\tag{A.2.9}
$$

where

$$
\boldsymbol{Z}_n = \begin{pmatrix} Z_{n,\boldsymbol{\theta}(Y_{(1)})} \\ \cdots \\ Z_{n,\boldsymbol{\theta}(Y_{(n)})} \end{pmatrix}
$$

and

$$
Z_{n,\boldsymbol{\theta}(Y_{(b)})} = \frac{1}{\sqrt{n}}\begin{pmatrix} \lambda_0 Q_{00}(Y_{(b)})(\mathbb{F}_{00}(Y_{(b)}) - F_{00}(Y_{(b)})) \\ (1-\lambda_0)Q_{00}(Y_{(b)})(\mathbb{F}_{00}(Y_{(b)}) - F_{00}(Y_{(b)})) + Q_{10}(Y_{(b)})(\mathbb{F}_{10}(Y_{(b)}) - F_{10}(Y_{(b)})) \\ \lambda_1 Q_{11}(Y_{(b)})(\mathbb{F}_{11}(Y_{(b)}) - F_{11}(Y_{(b)})) \\ (1-\lambda_1)Q_{11}(Y_{(b)})(\mathbb{F}_{11}(Y_{(b)}) - F_{11}(Y_{(b)})) + Q_{01}(Y_{(b)})(\mathbb{F}_{01}(Y_{(b)}) - F_{01}(Y_{(b)})) \end{pmatrix}.
\tag{A.2.10}
$$

Proof. Recall the definitions of $\mathbb{M}_n$ and $\mathbb{M}$ from (2.3.6) and (2.3.7). Then

$$
\begin{aligned}
&(\mathbb{M}_n - \mathbb{M})(\tilde{\boldsymbol{\theta}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F})\\
&= \frac{1}{n}\sum_{u,v\in\{0,1\}}\sum_{b\in I_\kappa}\frac{n_{uv}}{n}\left\{\mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})\right\}\left(\log\frac{\tilde{\theta}_{uv}(Y_{(b)})}{1-\tilde{\theta}_{uv}(Y_{(b)})} - \log\frac{F_{uv}(Y_{(b)})}{1-F_{uv}(Y_{(b)})}\right)
\end{aligned}
\tag{A.2.11}
$$

The log subtraction part in the parenthesis of equation (A.2.11) is

$$\log \frac{\tilde{\theta}_{uv}(Y_{(b)})}{1-\tilde{\theta}_{uv}(Y_{(b)})} - \log \frac{F_{uv}(Y_{(b)})}{1-F_{uv}(Y_{(b)})} = \log \frac{\tilde{\theta}_{uv}(Y_{(b)})}{F_{uv}(Y_{(b)})} - \log \frac{1-\tilde{\theta}_{uv}(Y_{(b)})}{1-F_{uv}(Y_{(b)})}$$
$$= \frac{\tilde{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})}{F_{uv}(Y_{(b)})(1-F_{uv}(Y_{(b)}))} + R_{uv}(Y_{(b)}), \tag{A.2.12}$$

where

$$R_{uv}(Y_{(b)}) = \frac{1}{2}(\tilde{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}))^2 \left(\frac{1}{(1-\gamma_{uv}(Y_{(b)}))^2} - \frac{1}{\gamma_{uv}(Y_{(b)})^2} \right),$$

where $\gamma_{uv}(Y_{(b)}) \in [F_{uv}(Y_{(b)}) \wedge \tilde{\theta}_{uv}(Y_{(b)}), F_{uv}(Y_{(b)}) \vee \tilde{\theta}_{uv}(Y_{(b)})]$.

Next, note that $\gamma_{uv}(Y_{(b)}) \geq \tilde{\theta}_{uv}(Y_{(b)}) \geq \tilde{\theta}_{uv}(Y_{(\lceil n\kappa \rceil)}) = F_{uv}(Y_{(\lceil n\kappa \rceil)}) + o_P(1)$, for $u, v \in \{0,1\}$. Moreover, by Observation A.2.2 $F_{uv}(Y_{(b)}) \geq F_{uv}(Y_{(\lceil n\kappa \rceil)}) = H_{uv}^{-1}(\kappa) + o_P(1)$. This implies that there exists a constant $0 < \delta(\kappa) < 1$, such that $\omega_{uv}(Y_{(b)}) \in [\delta(\kappa), 1-\delta(\kappa)]$ with high probability. Therefore, with a $O_P(1)$ term depending on the constant κ

$$\frac{1}{n} \sum_{b \in I_\kappa} |R_{uv}(Y_{(b)})| = O_P(1) \cdot \frac{1}{n} \sum_{b \in I_\kappa} (\tilde{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}))^2 = O_P(1/n), \tag{A.2.13}$$

by assumption. Then, using $||\mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})||_\infty = O_P(n^{-1/2})$ and (A.2.13),

$$\frac{1}{n} \sum_{u,v \in \{0,1\}} \frac{n_{uv}}{n} \sum_{b \in I_\kappa} (\mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})) R_{uv}(Y_{(b)}) = O_P(n^{-3/2}).$$

Combining this with (A.2.11) and (A.2.12), gives

$$(\mathbb{M}_n - \mathbb{M})(\tilde{\boldsymbol{\theta}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F})$$
$$= \frac{1}{n} \sum_{u,v \in \{0,1\}} \sum_{b \in I_\kappa} \frac{n_{uv}}{n} \frac{(\mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)}))}{F_{uv}(Y_{(b)})(1-F_{uv}(Y_{(b)}))} (\tilde{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})) + O_P(n^{-3/2}), \tag{A.2.14}$$

Then the (b)-th term of the sum in (A.2.12) above can be represented in terms of

$$\tilde{\theta}_{co}^{(0)}(Y_{(b)}) - F_{co}^{(0)}(Y_{(b)}), \quad \tilde{\theta}_{nt}(Y_{(b)}) - F_{nt}(Y_{(b)}), \quad \tilde{\theta}_{co}^{(1)}(Y_{(b)}) - F_{co}^{(1)}(Y_{(b)}), \quad \tilde{\theta}_{at}(Y_{(b)}) - F_{at}(Y_{(b)}),$$

as follows:

$$
\begin{aligned}
&\left\{\lambda_0 Q_{00}(Y_{(b)})(\overline{F}_{00}(Y_{(b)}) - F_{00}(Y_{(b)}))\right\}(\tilde{\theta}^{(0)}_{co}(Y_{(b)}) - F^{(0)}_{co}(Y_{(b)})) \\
&+\left\{(1-\lambda_0)Q_{00}(Y_{(b)})(\hat{F}_{00}(Y_{(b)}) - F_{00}(Y_{(b)})) + Q_{10}(Y_{(b)})(\hat{F}_{10}(Y_{(b)}) - F_{10}(Y_{(b)}))\right\}(\tilde{\theta}_{nt}(Y_{(b)}) - F_{nt}(Y_{(b)})) \\
&+\left\{Q_{11}(Y_{(b)})(\hat{F}_{11}(Y_{(b)}) - F_{11}(Y_{(b)}))\right\}(\tilde{\theta}^{(1)}_{co}(Y_{(b)}) - F^{(1)}_{co}(Y_{(b)})) \\
&+\left\{(1-\lambda_1)Q_{11}(Y_{(b)})(\hat{F}_{11}(Y_{(b)}) - F_{11}(Y_{(b)})) + Q_{01}(Y_{(b)})(\hat{F}_{01}(Y_{(b)}) - F_{01}(Y_{(b)}))\right\}(\tilde{\theta}_{at}(Y_{(b)}) - F_{at}(Y_{(b)})),
\end{aligned}
$$

where Q_{uv} is defined in (A.2.8)

Then from (A.2.14) and the definition of the vector $\boldsymbol{Z}_n(\cdot)$ in (A.2.10),

$$
(\mathbb{M}_n - \mathbb{M})(\tilde{\boldsymbol{\theta}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F}) = \frac{1}{\sqrt{n}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{F})^T \boldsymbol{Z}_n + O_P(n^{-\frac{3}{2}}). \tag{A.2.15}
$$

Finally, by a second order Taylor expansion of $\mathbb{M}(\tilde{\boldsymbol{\theta}}) - \mathbb{M}(\boldsymbol{F})$ around $((\boldsymbol{F}(Y_{(b)}))_{b\in I_\kappa}, \boldsymbol{\lambda})$,

$$
\mathbb{M}(\tilde{\boldsymbol{\theta}}) - \mathbb{M}(\boldsymbol{F}) = \frac{1}{2}\cdot(\tilde{\boldsymbol{\theta}} - \boldsymbol{F})^T \boldsymbol{V}(\tilde{\boldsymbol{\theta}} - \boldsymbol{F}) + O_P(n^{-\frac{3}{2}}), \tag{A.2.16}
$$

which, together with (A.2.15), implies the result in (A.2.16) follows since the gradient of $\mathbb{M}(\boldsymbol{F})$ at the point $((\boldsymbol{F}(Y_{(b)}))_{b\in I_\kappa})$ is zero. □

Lemma A.2.5. *Let* $(\hat{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\theta}\in\vartheta_+} \mathbb{M}_n(\boldsymbol{\theta})$ *be the BL estimate. Then the following holds:*

(a) $\frac{1}{n}\sum_{b\in I_\kappa} ||\hat{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 = O_P(1/n)$ *a.*

(b) For every finite set of indices $J \subseteq I_\kappa$, $(\sqrt{n}|\hat{\boldsymbol{\theta}}_n(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})|)_{b\in J} = O_P(1)$.

Proof. Since $\|\boldsymbol{\theta}-\boldsymbol{F}\|^2 = \frac{1}{n}\sum_{b\in I_\kappa}\|\boldsymbol{\theta}-\boldsymbol{F}\|_2^2$, define $\bar{B}(\boldsymbol{F},\delta) := \{\boldsymbol{\theta}\in\vartheta_+ : \delta/2 < \|\boldsymbol{\theta}-\boldsymbol{F}\| < \delta\}$. From the proof of Lemma A.2.3,

$$
\begin{aligned}
\max_{\boldsymbol{\theta}\in\bar{B}(\boldsymbol{F},\delta)} \mathbb{M}(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{F}) &\lesssim -\frac{1}{n}\sum_{b\in I_\kappa}\|\boldsymbol{\theta} - \boldsymbol{F}\|_2^2 \\
&= -\|\boldsymbol{\theta} - \boldsymbol{F}\|^2 < -\delta^2
\end{aligned} \tag{A.2.17}
$$

Now, by a first order Taylor expansion and arguments same as before, it follows that for

$\hat{\boldsymbol{\theta}} \in \bar{B}(\boldsymbol{F}, \delta)$

$$
\begin{aligned}
&|(\mathbb{M}_n - \mathbb{M})(\hat{\boldsymbol{\theta}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F})| \\
&\lesssim \frac{1}{n} \sum_{u,v \in \{0,1\}} \frac{n_{uv}}{n} \sum_{b \in I_\kappa} \left\{ |\mathbb{F}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})| |\hat{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})| \right\} \\
&\leq O_P\left(\frac{1}{\sqrt{n}}\right) \cdot \frac{1}{n} \sum_{u,v \in \{0,1\}} \sum_{b \in I_\kappa} |\hat{\theta}_{uv}(Y_{(b)}) - F_{uv}(Y_{(b)})| \\
&\leq O_P\left(\sqrt{\frac{\delta}{n}}\right), \qquad \text{(A.2.18)}
\end{aligned}
$$

where the last step uses the Cauchy-Schwarz inequality. Combining (A.2.17) and (A.2.18), and invoking (Van der Vaart and Wellner, 1996, Theorem 3.4.1), part (a) follows.

Define $\bar{D}(\boldsymbol{F}, \delta) := \left\{ \boldsymbol{\theta} \in \vartheta_+ : \delta/2 < \frac{1}{|J|} \sum_{b \in J} ||\boldsymbol{\theta}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 < \delta \right\}$. Then, from the proof of Lemma A.2.3, it is easy to see that

$$
\max_{\boldsymbol{\theta} \in \bar{D}(\boldsymbol{F}, \delta)} \mathbb{M}(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{F}) \lesssim -\delta. \qquad \text{(A.2.19)}
$$

For $\hat{\boldsymbol{\theta}} \in \bar{D}(\boldsymbol{F}, \delta)$, similar to (A.2.18),

$$
\begin{aligned}
|(\mathbb{M}_n - \mathbb{M})(\hat{\boldsymbol{\theta}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F})| &\leq O_P\left(\frac{1}{\sqrt{n}}\right) \cdot \sum_{u,v \in \{0,1\}} \sum_{b \in J} |\hat{\theta}_{uv} Y_{(b)}) - F_{uv}(Y_{(b)})| \\
&\leq O_P\left(\sqrt{\frac{\delta}{n}}\right). \qquad \text{(A.2.20)}
\end{aligned}
$$

Finally, as before, by (Van der Vaart and Wellner, 1996, Theorem 3.4.1), $\frac{1}{|J|} \sum_{b \in J} ||\hat{\boldsymbol{\theta}}(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})||_2^2 = O_P(1/n)$, which implies $(\sqrt{n}|\hat{\boldsymbol{\theta}}_n(Y_{(b)}) - \boldsymbol{F}(Y_{(b)})|)_{b \in J} = O_P(1)$. □

Recall $\boldsymbol{Z}_n$ and $\boldsymbol{V}_n$ from (A.2.10) and (A.2.6), respectively. Now, substitute $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$, the BL estimate, in (A.2.9). From Lemma A.2.4 and Lemma A.2.5,

$$
\mathbb{M}_n(\hat{\boldsymbol{\theta}}) - \mathbb{M}_n(\boldsymbol{F}) = \frac{1}{\sqrt{n}} (\hat{\boldsymbol{\theta}} - \boldsymbol{F})^T \boldsymbol{Z}_n + \frac{1}{2} (\hat{\boldsymbol{\theta}} - \boldsymbol{F})^T \boldsymbol{V}_n (\hat{\boldsymbol{\theta}} - \boldsymbol{F}) + O_P(n^{-\frac{3}{2}}). \qquad \text{(A.2.21)}
$$

Then substitute $\tilde{\boldsymbol{\theta}} = \boldsymbol{F} - n^{-\frac{1}{2}} \boldsymbol{V}_n^{-1} \boldsymbol{Z}_n$ in (A.2.15), where $\boldsymbol{Z}_n = (\boldsymbol{Z}_n(Y_{(b)})')'_{b \in I_\kappa}$ and $\boldsymbol{V}_n = \text{diag}(\boldsymbol{V}_n(Y_{(b)}))_{b \in I_\kappa}$. By using the matrix $\boldsymbol{V}_n^{-1}$ as computed in (A.2.7), a direct multiplication

shows that

$$-\boldsymbol{V}_n^{-1}Z_n = \begin{pmatrix} \vdots \\ \frac{\mathbb{G}_{00}(Y_{(b)})-(1-\lambda_0)\mathbb{G}_{10}(Y_{(b)})}{\lambda_0} \\ \mathbb{G}_{10}(Y_{(b)}) \\ \frac{\mathbb{G}_{11}(Y_{(b)})-(1-\lambda_1)\mathbb{G}_{01}(Y_{(b)})}{\lambda_1} \\ \mathbb{G}_{01}(Y_{(b)}) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \sqrt{n}\left(\frac{\mathbb{F}_{00}(Y_{(b)})-(1-\lambda_0)\mathbb{F}_{10}(Y_{(b)})}{\lambda_0} - F_{co}^{(0)}(Y_{(b)})\right) + O_P(1) \\ \sqrt{n}(\mathbb{F}_{10} - F_{nt})(Y_{(b)}) + O_P(1) \\ \sqrt{n}\left(\frac{\mathbb{F}_{11}(Y_{(b)})-(1-\lambda_1)\mathbb{F}_{01}(Y_{(b)})}{\lambda_1} - F_{co}^{(1)}(Y_{(b)})\right) + O_P(1) \\ \sqrt{n}(\mathbb{F}_{01} - F_{at})(Y_{(b)}) + O_P(1) \\ \vdots \end{pmatrix}$$

where $\mathbb{G}_{uv}(t) = \sqrt{n}(\mathbb{F}_{uv}(t) - F_{uv}(t))$. We let

$$\begin{aligned} \breve{F}_{co}^{(0)}(Y_{(b)}) &= \frac{\mathbb{F}_{00}(Y_{(b)}) - (1-\lambda_0)\mathbb{F}_{10}(Y_{(b)})}{\lambda_0} \\ \breve{F}_{nt}(Y_{(b)}) &= \mathbb{F}_{10}(Y_{(b)}) \\ \breve{F}_{co}^{(1)}(Y_{(b)}) &= \frac{\mathbb{F}_{11}(Y_{(b)}) - (1-\lambda_1)\mathbb{F}_{01}(Y_{(b)})}{\lambda_1} \\ \breve{F}_{at}(Y_{(b)}) &= \mathbb{F}_{01}(Y_{(b)}). \end{aligned} \tag{A.2.22}$$

Then, the estimators $\breve{F}_{co}^{(0)}, \breve{F}_{nt}, \breve{F}_{(1)}, \breve{F}_{at}$ are the MOM estimators discussed in Abadie (2002). Furthermore, the term $-\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n$ is further simplified as

$$-\boldsymbol{V}_n^{-1}Z_n = \begin{pmatrix} \vdots \\ \sqrt{n}(\breve{F}_{co}^{(0)} - F_{co}^{(0)})(Y_{(b)}) + O_P(1) \\ \sqrt{n}(\breve{F}_{nt} - F_{nt})(Y_{(b)}) + O_P(1) \\ \sqrt{n}(\breve{F}_{co}^{(1)} - F_{co}^{(1)})(Y_{(b)}) + O_P(1) \\ \sqrt{n}(\breve{F}_{at} - F_{at})(Y_{(b)}) + O_P(1) \\ \vdots \end{pmatrix} = \sqrt{n}(\breve{\boldsymbol{F}} - \boldsymbol{F} + O_P(n^{-1/2})) \tag{A.2.23}$$

Therefore, $\frac{1}{n}\sum_{b\in I_\kappa} ||\frac{1}{\sqrt{n}}\boldsymbol{V}_n^{-1}(Y_{(b)})\boldsymbol{Z}_n(Y_{(b)})||_2^2 = O_P(1/n)$, since $\sup_t |\mathbb{G}_{uv}(t)| = O_P(1/\sqrt{n})$. This implies,

$$\mathbb{M}_n(\boldsymbol{F} - n^{-\frac{1}{2}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n) - \mathbb{M}_n(\boldsymbol{F}) = -\frac{1}{2n}\boldsymbol{Z}_n'\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n + O_P(n^{-\frac{3}{2}}). \tag{A.2.24}$$

Subtracting (A.2.24) from (A.2.21) gives,

$$\begin{aligned}
&\mathbb{M}_n(\hat{\boldsymbol{\theta}}) - \mathbb{M}_n(\boldsymbol{F} - \frac{1}{\sqrt{n}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n) \\
&= \frac{1}{\sqrt{n}}(\hat{\boldsymbol{\theta}} - \boldsymbol{F})^T\boldsymbol{Z}_n + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{F})^T\boldsymbol{V}_n(\hat{\boldsymbol{\theta}} - \boldsymbol{F}) + \frac{1}{2n}\boldsymbol{Z}_n'\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n + O_P(n^{-\frac{3}{2}}) \\
&= \frac{1}{2}\left((\hat{\boldsymbol{\theta}} - \boldsymbol{F}) + \frac{1}{\sqrt{n}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n\right)^T\boldsymbol{V}_n\left((\hat{\boldsymbol{\theta}} - \boldsymbol{F}) + \frac{1}{\sqrt{n}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n\right) + O_P(n^{-\frac{3}{2}}) \\
&= \frac{1}{2}(\hat{\boldsymbol{\theta}} - \breve{\boldsymbol{F}})^T\boldsymbol{V}_n(\hat{\boldsymbol{\theta}} - \breve{\boldsymbol{F}}) + O_P(n^{-\frac{3}{2}}), \qquad \text{(A.2.25)}
\end{aligned}$$

where $\breve{\boldsymbol{F}}(\cdot)$ is defined in (A.2.22).

Lemma A.2.6. *The BL estimate $\hat{\boldsymbol{\theta}}$ satisfies*

$$\frac{1}{n}\sum_{b\in I_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{\theta}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})\}||_2^2 = O_P(1/\sqrt{n}).$$

Proof. For $b \in [n]$, denote by $\boldsymbol{x}_b = \sqrt{n}\{\hat{\boldsymbol{\theta}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})\}$ and $\boldsymbol{A}_b = -\boldsymbol{V}_n(Y_{(b)})$. Then (A.2.25) implies,

$$\frac{1}{n}\sum_{b\in I_\kappa} \boldsymbol{x}_b'\boldsymbol{A}_b\boldsymbol{x}_b = O_P(1/\sqrt{n}), \qquad \text{(A.2.26)}$$

since $\mathbb{M}_n(\hat{\boldsymbol{\theta}}) - \mathbb{M}_n(\boldsymbol{\theta}_0 - \frac{1}{\sqrt{n}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n) \geq 0$.

Note that $\boldsymbol{A}_b$ is positive definite, and denote by $||\boldsymbol{A}_b^{-1}||_\infty$ the maximum eigenvalue of $\boldsymbol{A}_b^{-1}$. Let $\boldsymbol{A}_b = \sum_{j=1}^4 \lambda_{bj}\boldsymbol{p}_{bj}\boldsymbol{p}_{bj}'$ be the spectral decomposition of $\boldsymbol{A}_b$, where $\lambda_{b1} \leq \lambda_{b2} \leq \lambda_{b3} \leq \lambda_{b4}$ are the eigenvalues of $\boldsymbol{A}_b$. This implies

$$\boldsymbol{x}_b'\boldsymbol{A}_b\boldsymbol{x}_b = \sum_{j=1}^4 \lambda_{bj}|\boldsymbol{p}_{bj}'\boldsymbol{x}_b|^2 \geq \lambda_{b1}\sum_{j=1}^3 |\boldsymbol{p}_{bj}'\boldsymbol{x}_b|^2 = \lambda_{b1}||\boldsymbol{x}_b||_2^2 = \frac{1}{||\boldsymbol{A}_b^{-1}||_\infty}||\boldsymbol{x}_b||_2^2. \qquad \text{(A.2.27)}$$

Next, recall $Q_{uv}(\cdot)$ from (A.2.8), and note that

$$\frac{4n_{uv}}{n} \leq Q_{uv}(Y_{(b)}) \leq \frac{1}{F_{uv}(Y_{\lceil n\kappa\rceil})(1 - F_{uv}(Y_{\lceil n(1-\kappa)\rceil}))},$$

where the upper bound uses the monotonicity of the distribution function and lower bound uses $x(1-x) \leq 1/4$. Now, there exists $\delta(\kappa)$ such that $F_{uv}(Y_{\lceil n\kappa\rceil}) \in [\delta(\kappa), 1]$ and $F_{uv}(Y_{\lceil n(1-\kappa)\rceil}) \in [0, 1-\delta(\kappa)]$ with high probability. Finally, since n_{uv}/n converges in probability to $\eta_{uv} > 0$, there exists $0 < c(\kappa) < C(\kappa) < \infty$ such that

$$\mathbb{P}(c(\kappa) < \sup_{b\in I_\kappa} Q_{uv}(Y_{(b)}) < C(\kappa)) \to 1.$$

This implies, $\sup_{b\in I_\kappa} ||\boldsymbol{A}_b^{-1}||_\infty \leq \sup_{b\in I_\kappa} \max_i ||\boldsymbol{A}_b^{-1}\boldsymbol{e}_i||_2^2 < M(\kappa)$, with high probability, for

some $M(\kappa) < \infty$, that is, $\sup_{b\in I_\kappa} ||\boldsymbol{A}_b^{-1}||_\infty = O_P(1)$. Then, from (A.2.27)

$$\begin{aligned}\frac{1}{n}\sum_{b\in I_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{\theta}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})\}||_2^2 = \frac{1}{n}\sum_{b\in I_\kappa} ||\boldsymbol{x}_b||_2^2 &\leq \sup_{b\in I_\kappa} ||\boldsymbol{A}_b^{-1}||_\infty \left(\frac{1}{n}\sum_{b\in I_\kappa} \boldsymbol{x}_b'\boldsymbol{A}_b\boldsymbol{x}_b\right) \\ &= O_P(1)\left(\frac{1}{n}\sum_{b\in I_\kappa} \boldsymbol{x}_b'\boldsymbol{A}_b\boldsymbol{x}_b\right) \\ &= O_P(1/\sqrt{n}),\end{aligned}$$

where the last step uses (A.2.26). □

A.2.3. Proof of Theorem 2.4.1

We define the MBL estimator under the general alternative as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\vartheta_+} \ell(\boldsymbol{\theta})$ and the MBL estimator under the null as $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\theta}\in\vartheta_{+,0}} \ell(\boldsymbol{\theta})$. Recall that $\hat{\boldsymbol{\theta}}(t) = (\hat{\theta}_{co}^{(0)}(t), \hat{\theta}_{nt}(t), \hat{\theta}_{co}^{(1)}(t), \hat{\theta}_{at}(t))$, and $\hat{\boldsymbol{\psi}}(t) = (\hat{\psi}_{co}(t), \hat{\psi}_{nt}(t), \hat{\psi}_{at}(t))$. Introduce the following functions,

$$\begin{aligned}\hat{\theta}_{00}(t) = \lambda_0\hat{\theta}_{co}^{(0)}(t) + (1-\lambda_0)\hat{\theta}_{nt}(t), \qquad & \hat{\psi}_{00}(t) = \lambda_0\hat{\psi}_{co}(t) + (1-\lambda_0)\hat{\psi}_{nt}(t) \\ \hat{\theta}_{01}(t) = \hat{\theta}_{at}(t), \qquad & \hat{\psi}_{01}(t) = \hat{\psi}_{at}(t) \\ \hat{\theta}_{10}(t) = \hat{\theta}_{nt}(t), \qquad & \hat{\psi}_{10}(t) = \hat{\psi}_{nt}(t) \\ \hat{\theta}_{11}(t) = \lambda_1\hat{\theta}_{co}^{(1)}(t) + (1-\lambda_1)\hat{\theta}_{at}(t), \qquad & \hat{\psi}_{11}(t) = \lambda_1\hat{\psi}_{co}(t) + (1-\lambda_1)\hat{\psi}_{at}(t).\end{aligned}$$

Lemma A.2.7. *Let $\hat{\theta}_{00}, \hat{\theta}_{01}, \hat{\theta}_{10}, \hat{\theta}_{11}$, and $\hat{\psi}_{00}, \hat{\psi}_{01}, \hat{\psi}_{10}, \hat{\psi}_{11}$ be as defined above. Then the BLRT statistic satisfies*

$$T_n = \frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa} n_{uv}\left\{\frac{(\hat{\psi}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^2}{\mathbb{F}_{uv}(Y_{(b)})(1-\mathbb{F}_{uv}(Y_{(b)}))} - \frac{(\hat{\theta}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^2}{\mathbb{F}_{uv}(Y_{(b)})(1-\mathbb{F}_{uv}(Y_{(b)}))}\right\} + O_P(1/\sqrt{n}). \tag{A.2.28}$$

Proof. Then, the difference of the log-BL under the alternative and the null can written as follows:

$$\mathbb{M}_n(\hat{\boldsymbol{\theta}}) - \mathbb{M}_n(\hat{\boldsymbol{\psi}}) = \frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa}(T_{uv}(Y_{(b)})|\hat{\boldsymbol{\theta}}) - T_{uv}(Y_{(b)})|\hat{\boldsymbol{\psi}}), \tag{A.2.29}$$

where

$$\begin{aligned}T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) =& \frac{n_{uv}}{n}\left\{\mathbb{F}_{uv}(Y_{(b)})\log\hat{\theta}_{uv}(Y_{(b)}) + (1-\mathbb{F}_{uv}(Y_{(b)}))\log(1-\hat{\theta}_{uv}(Y_{(b)}))\right\} \\ T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}) =& \frac{n_{uv}}{n}\left\{\mathbb{F}_{uv}(Y_{(b)})\log\hat{\psi}_{uv}(Y_{(b)}) + (1-\mathbb{F}_{uv}(Y_{(b)}))\log(1-\hat{\psi}_{uv}(Y_{(b)}))\right\}.\end{aligned}$$

Now, adding and subtracting the negative binary entropy $J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)}))$, it follows that

$$T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}) = \left\{T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)}))\right\} - \left\{T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}) - J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)}))\right\}. \tag{A.2.30}$$

Next, note that

$$\begin{aligned} &T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)})) \\ &= \frac{n_{uv}}{n}\left\{\mathbb{F}_{uv}(Y_{(b)}) \log \frac{\hat{\theta}_{uv}(Y_{(b)})}{\mathbb{F}_{uv}(Y_{(b)})} + (1 - \mathbb{F}_{uv}(Y_{(b)})) \log \frac{1 - \hat{\theta}_{uv}(Y_{(b)})}{1 - \mathbb{F}_{uv}(Y_{(b)})}\right\} \\ &= \frac{n_{uv}}{n} \cdot \frac{1}{2} \cdot \frac{(\hat{\theta}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^2}{\mathbb{F}_{uv}(Y_{(b)})(1 - \mathbb{F}_{uv}(Y_{(b)}))} + R_{uv}^{(b)}, \end{aligned} \tag{A.2.31}$$

where

$$R_{uv}^{(b)} = \frac{n_{uv}}{n} \cdot \frac{(\hat{\theta}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^3}{6}\left\{\frac{\mathbb{F}_{uv}(Y_{(b)})}{(\omega_{uv}(Y_{(b)}))^3} - \frac{1 - \mathbb{F}_{uv}(Y_{(b)})}{(1 - \omega_{uv}(Y_{(b)}))^3}\right\},$$

where $\omega_{uv}(Y_{(b)}) \in [\mathbb{F}_{uv}(Y_{(b)}) \wedge \hat{\theta}_{uv}(Y_{(b)}), \hat{\theta}_{uv}(Y_{(b)}) \vee \mathbb{F}_{uv}(Y_{(b)})]$.

Next, note that $\omega_{uv}(Y_{(b)}) \geq \hat{\theta}_{uv}(Y_{(b)}) \geq \hat{\theta}_{uv}(Y_{(\lceil n\kappa \rceil)}) = F_{uv}(Y_{(\lceil n\kappa \rceil)}) + o_P(1)$, for $u, v \in \{0, 1\}$. Moreover, $\mathbb{F}_{uv}(Y_{(b)}) \geq \mathbb{F}_{uv}(Y_{(\lceil n\kappa \rceil)}) = F_{uv}(Y_{(\lceil n\kappa \rceil)}) + o_P(1)$. Finally, since $F_{uv}(Y_{(\lceil n\kappa \rceil)}) = G_{uv}^{-1}(\kappa) + o_P(1)$ (Notation change). This implies that there exists a constant $0 < \delta(\kappa) < 1$, such that $\omega_{uv}(Y_{(b)}) \in [\delta(\kappa), 1 - \delta(\kappa)]$ with high probability. Then, since κ is a constant,

$$\begin{aligned} \sum_{b \in I_\kappa} |R_{uv}^{(b)}| &\leq O_P(1) \sum_{b \in I_\kappa} |\hat{\theta}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)})|^3 \\ &\leq O_P(1) \sum_{b \in I_\kappa} |F_{uv}(Y_{(b)}) - \hat{\theta}_{uv}(Y_{(b)})|^3 + O_P(1/\sqrt{n}). \end{aligned} \tag{A.2.32}$$

Now, note that

$$\begin{aligned} &\frac{1}{n^{\frac{3}{2}}} \sum_{b \in I_\kappa} |\sqrt{n}[F_{uv}(Y_{(b)}) - \hat{\theta}_{uv}(Y_{(b)})]|^3 \\ &\lesssim \frac{1}{n^{\frac{3}{2}}} \sum_{b \in I_\kappa} |\sqrt{n}\{F_{uv}(Y_{(b)}) - \hat{\theta}_{uv}(Y_{(b)})\} + \boldsymbol{V}_n^{-1}(Y_{(b)})\boldsymbol{Z}_n(Y_{(b)})|^3 + \frac{1}{n^{\frac{3}{2}}} \sum_{b \in I_\kappa} |\boldsymbol{V}_n^{-1}(Y_{(b)})\boldsymbol{Z}_n(Y_{(b)})|^3 \\ &\leq \left[\frac{1}{n} \sum_{b \in I_\kappa} |\sqrt{n}\{F_{uv}(Y_{(b)}) - \hat{\theta}_{uv}(Y_{(b)})\} + \boldsymbol{V}_n^{-1}(Y_{(b)})\boldsymbol{Z}_n(Y_{(b)})|^2\right]^{\frac{3}{2}} + O_P(1/\sqrt{n}) = o_P(1), \end{aligned} \tag{A.2.33}$$

using Theorem 2.3.2 and that $\max_{b \in I_\kappa} |V^{-1}(Y_{(b)})\boldsymbol{Z}(Y_{(b)})| = O_P(1)$. Finally, combining

(A.2.32) and (A.2.33), it follows that $\sum_{b\in I_\kappa}|R_{uv}^{(b)}| = o_P(1)$. This implies, by (A.2.31),

$$T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)})) = \frac{n_{uv}}{n}\cdot\frac{1}{2}\cdot\frac{(\hat{\theta}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^2}{\mathbb{F}_{uv}(Y_{(b)})(1-\mathbb{F}_{uv}(Y_{(b)}))} + o_P(1). \tag{A.2.34}$$

Similarly,

$$T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}) - J(\mathbb{F}_{uv}(Y_{(b)}), \mathbb{F}_{uv}(Y_{(b)})) = \frac{n_{uv}}{n}\cdot\frac{1}{2}\cdot\frac{(\hat{\psi}_{uv}(Y_{(b)}) - \mathbb{F}_{uv}(Y_{(b)}))^2}{\mathbb{F}_{uv}(Y_{(b)})(1-\mathbb{F}_{uv}(Y_{(b)}))} + o_P(1). \tag{A.2.35}$$

Combining (A.2.34) and (A.2.35) with (A.2.29) and (A.2.30) the result follows. □

Proposition A.2.1. *Under the null H_0,*

$$\frac{1}{n}\sum_{b\in I_\kappa}\left\|\begin{pmatrix}\sqrt{n}[\hat{\theta}_{co}(Y_{(b)}) - \hat{\tau}_{co}(Y_{(b)})]\\ \sqrt{n}[\hat{\theta}_{nt}(Y_{(b)}) - \hat{\tau}_{nt}(Y_{(b)})]\\ \sqrt{n}[\hat{\theta}_{co}(Y_{(b)}) - \hat{\tau}_{at}(Y_{(b)})]\end{pmatrix}\right\|_2^2 = o_P(1)$$

where

$$\begin{aligned}
\hat{\tau}_{co}(t) &= \frac{(C_{01}(t)+C_{11}(t))\left(\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0}\right) + (C_{10}(t)+C_{00}(t))\left(\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}\right)}{C(t)}\\
\hat{\tau}_{nt}(t) &= \mathbb{F}_{10}(t) + \frac{\frac{\lambda_0}{1-\lambda_0}C_{10}(t)}{C(t)}\left\{\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0} - \frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}\right\}\\
\hat{\tau}_{at}(t) &= \mathbb{F}_{01}(t) + \frac{\frac{\lambda_1}{1-\lambda_1}C_{01}(t)}{C(t)}\left\{\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1} - \frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0}\right\}.
\end{aligned} \tag{A.2.36}$$

and

$$\begin{aligned}
C_{00}(t) &= \frac{F_{00}(t)(1-F_{00}(t))}{\eta_{00}\lambda_1^2}\\
C_{01}(t) &= \frac{F_{01}(t)(1-F_{01}(t))}{\eta_{01}\lambda_0^2(1-\lambda_1)^2}\\
C_{10}(t) &= \frac{F_{10}(t)(1-F_{10}(t))}{\eta_{10}(1-\lambda_0)^2\lambda_1^2}\\
C_{11}(t) &= \frac{F_{11}(t)(1-F_{11}(t))}{\eta_{11}\lambda_0^2}\\
C(t) &= C_{00}(t)+C_{01}(t)+C_{10}(t)+C_{11}(t)
\end{aligned}$$

Proof. Similar to the proof in Theorem 2.3.2. □

From Theorem 2.3.2, it follows that

$$\frac{1}{n}\sum_{u,v\in\{0,1\}}\sum_{b\in I_\kappa}\Big(\sqrt{n}[\tilde{\theta}_{uv}(Y_{(b)})-\mathbb{F}_{uv}(Y_{(b)})]\Big)^2=o_P(1).$$

Next, denote

$$B_0(t)=\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0},\quad B_1(t)=\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}.$$

Then

$$\begin{aligned}
\hat{\tau}_{00}(t)&=\mathbb{F}_{00}(t)+\frac{C_{01}(t)\lambda_0}{C(t)}\Big(\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}-\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0}\Big)+o_P(n^{-1/2})\\
&=\hat{F}_{00}(t)+\frac{C_{00}(t)\lambda_0}{C(t)}\Big(B_1(t)-B_0(t)\Big)\\
\hat{\tau}_{01}(t)&=\mathbb{F}_{01}(t)+\frac{C_{01}(t)\frac{\lambda_1}{1-\lambda_1}}{C(t)}\Big(\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}-\frac{\mathbb{F}_{00}(y_b)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0}\Big)\\
&=\mathbb{F}_{01}(t)+\frac{C_{01}(t)\frac{\lambda_1}{1-\lambda_1}}{C(t)}\Big(B_1(t)-B_0(t)\Big)\\
\hat{\tau}_{10}(t)&=\mathbb{F}_{10}(t)+\frac{C_{10}(t)\frac{\lambda_0}{1-\lambda_0}}{C(t)}\Big(\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0}-\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}\Big)\\
&=\mathbb{F}_{10}(t)+\frac{C_{10}(t)\frac{\lambda_0}{1-\lambda_0}}{C(t)}\Big(B_0(t)-B_1(t)\Big)\\
\hat{\tau}_{11}(t)&=\mathbb{F}_{11}(t)+\frac{C_{11}(t)\lambda_1}{C(t)}\Big(\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}}{\lambda_0}-\frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}\Big)+o_P(n^{-1/2})\\
&=\mathbb{F}_{11}(t)+\frac{C_{11}(t)\lambda_1}{C(t)}\Big(B_0(t)-B_1(t)\Big).
\end{aligned}$$

and by Proposition A.2.1, it follows that

$$\frac{1}{n}\sum_{u,v\in\{0,1\}}\sum_{b\in I_\kappa}\Big(\sqrt{n}[\tilde{\psi}_{uv}(Y_{(b)})-\hat{\tau}_{uv}(Y_{(b)})]\Big)^2=o_P(1).$$

By putting these asymptotically equivalent estimators into equation (A.2.28), we get

$$
\begin{aligned}
T_n = & \frac{1}{n}\sum_{b\in I_\kappa}\left[\frac{\lambda_0 C_{00}(Y_{(b)})\sqrt{\hat{Q}_{00}(Y_{(b)})}}{C(Y_{(b)})}\sqrt{n}(B_1(Y_{(b)})-B_0(Y_{(b)}))\right]^2 \\
& + \frac{1}{n}\sum_{b\in I_\kappa}\left[\frac{\frac{\lambda_1}{1-\lambda_1} C_{01}(Y_{(b)})\sqrt{\hat{Q}_{01}(Y_{(b)})}}{C(Y_{(b)})}\sqrt{n}(B_1(Y_{(b)})-B_0(Y_{(b)}))\right]^2 \\
& + \frac{1}{n}\sum_{b\in I_\kappa}\left[\frac{\frac{\lambda_0}{1-\lambda_0} C_{10}(Y_{(b)})\sqrt{\hat{Q}_{10}(Y_{(b)})}}{C(Y_{(b)})}\sqrt{n}(B_1(Y_{(b)})-B_0(Y_{(b)}))\right]^2 \\
& + \frac{1}{n}\sum_{b\in I_\kappa}\left[\frac{\lambda_1 C_{11}(Y_{(b)})\sqrt{\hat{Q}_{11}(Y_{(b)})}}{C(Y_{(b)})}\sqrt{n}(B_1(Y_{(b)})-B_0(Y_{(b)}))\right]^2 + o_P(1). \qquad \text{(A.2.37)}
\end{aligned}
$$

Note that $\hat{Q}_{uv}(t) \xrightarrow{P} \frac{\eta_{uv}}{F_{uv}(t)(1-F_{uv}(t))} := Q_{uv}(t)$ in the supremum norm. In fact,

$$
\sup_{b\in I_\kappa}\left|\frac{\hat{Q}_{uv}(Y_{(b)})}{Q_{uv}(Y_{(b)})}-1\right| = o_P(1).
$$

Define

$$
W(Y_{(b)}) = \frac{1}{\lambda_0^2 Q_{00}(Y_{(b)})} + \frac{(1-\lambda_1)^2}{\lambda_1^2 Q_{01}(Y_{(b)})} + \frac{(1-\lambda_0)^2}{\lambda_0^2 Q_{10}(Y_{(b)})} + \frac{1}{\lambda_1^2 Q_{11}(Y_{(b)})}.
$$

We have

$$
\begin{aligned}
T_n &= \frac{1}{n}\sum_{b\in I_\kappa}\frac{\frac{1}{\lambda_0^2}\frac{\hat{Q}_{00}(Y_{(b)})}{Q_{00}(Y_{(b)})^2} + \frac{(1-\lambda_1)^2}{\lambda_1^2}\frac{\hat{Q}_{01}(Y_{(b)})}{Q_{01}(Y_{(b)})^2} + \frac{(1-\lambda_0)^2}{\lambda_0^2}\frac{\hat{Q}_{10}(Y_{(b)})}{Q_{10}(Y_{(b)})^2} + \frac{1}{\lambda_1^2}\frac{\hat{Q}_{11}(Y_{(b)})}{Q_{11}(Y_{(b)})^2}}{W(Y_{(b)})^2}(\sqrt{n}\{B_0(Y_{(b)})-B_1(Y_{(b)})\})^2 + o \\
&= \frac{1}{n}\sum_{b\in I_\kappa}\frac{(\sqrt{n}\{B_0(Y_{(b)})-B_1(Y_{(b)})\})^2}{W(Y_{(b)})} + o_P(1) \qquad \text{(A.2.38)}
\end{aligned}
$$

using the above estimates and the definition of $W(t)$. Note that $\mathbb{E}(B_0(Y_{(b)})-B_1(Y_{(b)})|Y_{(b)} = t) = 0$ and variance

$$
\begin{aligned}
n\,\mathrm{Var}(B_0(t)-B_1(t)) &= \mathrm{Var}\left(\frac{\mathbb{F}_{00}(t)-(1-\lambda_0)\mathbb{F}_{10}(t)}{\lambda_0} - \frac{\mathbb{F}_{11}(t)-(1-\lambda_1)\mathbb{F}_{01}(t)}{\lambda_1}\right) \\
&= \frac{1}{\lambda_0^2 Q_{00}(t)} + \frac{(1-\lambda_1)^2}{\lambda_1^2 Q_{01}(t)} + \frac{(1-\lambda_0)^2}{\lambda_0^2 Q_{10}(t)} + \frac{1}{\lambda_1^2 Q_{11}(t)} \\
&= W(t).
\end{aligned}
$$

Recall the definitions of $B_0(t)$ and $B_1(t)$. Then $B_0(t)-B_1(t) = \frac{1}{\lambda_0}\hat{G}_{00}-\frac{1-\lambda_0}{\lambda_0}\hat{G}_{01}+\frac{1-\lambda_1}{\lambda_1}\hat{G}_{10}-\frac{1}{\lambda_1}\hat{G}_{11} = \hat{G}$, where $\hat{G}_{uv} = \mathbb{F}_{uv} - F_{uv}$, for $u, v \in \{0,1\}$. Note that $\sup_{t\in\mathbb{R}}\left|\sqrt{n}\hat{G}_n(t) - G(t)\right| = o_P(1)$, where

$$G(t) = \left(\sqrt{\frac{\phi_c+\phi_n}{\phi_0}}B_{00}(F_{00}) - \sqrt{\frac{\phi_n}{\phi_1}}B_{10}(F_{10})\right) - \left(\sqrt{\frac{\phi_c+\phi_a}{\phi_1}}B_{11}(F_{11}) - \sqrt{\frac{\phi_a}{\phi_0}}B_{01}(F_{01})\right),$$

for independent Brownian bridges $B_{00}, B_{01}, B_{10}, B_{11}$. This implies,

$$\begin{aligned} T_n = \frac{1}{n}\sum_{b\in I_\kappa}\frac{\{\sqrt{n}\hat{G}_n(Y_{(b)})\}^2}{W(Y_{(b)})} + o_P(1) &\overset{D}{=} \frac{1}{n}\sum_{b\in I_\kappa}\frac{G(Y_{(b)})^2}{W(Y_{(b)})} + o_P(1) \\ &= \int_\kappa^{1-\kappa}\frac{G(H^{-1}(s))^2}{W(H^{-1}(s))}\mathrm{d}s + o_P(1) \\ &= \int_{H^{-1}(\kappa)}^{H^{-1}(1-\kappa)}\frac{G(t)^2}{\mathrm{Var}(G(t))}\mathrm{d}H(t) + o_P(1). \end{aligned} \tag{A.2.39}$$

A.3. EM-PAVA algorithm

A.3.1. The pool-adjacent-violator (PAV) algorithm

The maximization (2.3.5) is not easy to solve because the parameter space ϑ_+ includes the non-decreasing condition. We use the EM algorithm to achieve the maximum. To apply the EM algorithm, we need to find the complete data binomial likelihood. Given that the compliance class indicator $\boldsymbol{S}$ is known, the complete data binomial likelihood $L_B(\boldsymbol{\theta})$ is defined as

$$\begin{aligned} &L_B(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S},\boldsymbol{Y}) \\ &= \prod_{a=1}^n\prod_{b=1}^n \theta_{co}^{(0)}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=0, D_a=0, S_a=co\}}(1-\theta_{co}^{(0)}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=0, D_a=0, S_a=co\}} \\ &\times\theta_{nt}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=0, D_a=0, S_a=nt\}}(1-\theta_{nt}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=0, D_a=0, S_a=nt\}} \\ &\times\theta_{at}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=0, D_a=1, S_a=at\}}(1-\theta_{at}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=0, D_a=1, S_a=at\}} \\ &\times\theta_{nt}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=1, D_a=0, S_a=nt\}}(1-\theta_{nt}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=1, D_a=0, S_a=nt\}} \\ &\times\theta_{co}^{(1)}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=1, D_a=1, S_a=co\}}(1-\theta_{co}^{(1)}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=1, D_a=1, S_a=co\}} \\ &\times\theta_{at}(Y_b)^{\mathbf{1}\{Y_a\le Y_b, Z_a=1, D_a=1, S_a=at\}}(1-\theta_{at}(Y_b))^{\mathbf{1}\{Y_a> Y_b, Z_a=1, D_a=1, S_a=at\}}. \end{aligned}$$

and the log complete data log-likelihood is defined as $\ell_B(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S},\boldsymbol{Y}) = \log L_B(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S},\boldsymbol{Y})$.

We use the following algorithm to solve the maximization:

1. Set an initial value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^{(0)}$.

2. Compute the expected value of the complete data log-likelihood, with respect to the

conditional distribution of $\boldsymbol{S}$ given observed data $(\boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{Y})$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = \mathbb{E}_{\boldsymbol{S}|\boldsymbol{Z},\boldsymbol{D},\boldsymbol{Y}}[\ell_B(\boldsymbol{\theta})].$$

3. Maximize the quantity $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$:

$$\boldsymbol{\theta}^{(1)}_{temp} = \arg\max_{\boldsymbol{\theta}\in\vartheta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$$

   However, $\boldsymbol{\theta}^{(1)}_{temp}$ is not in the parameter space ϑ_+ that is our interest. The maximization of this step can be easily solved.

4. Apply the pool-adjacent-violator algorithm (PAVA) to the temporarily obtained parameter $\boldsymbol{\theta}^{(1)}_{temp}$: $\boldsymbol{\theta}^{(1)} = f_{PAVA}(\boldsymbol{\theta}^{(1)}_{temp})$. The algorithm f_{PAVA} is illustrated in the following section.

5. Iterate the Step 2, Step 3 and Step 4 until the tolerance is achieved.

The algorithm implicitly assumes that the maximization $\arg\max_{\boldsymbol{\theta}\in\vartheta_+} \ell_B(\boldsymbol{\theta})$ is solved by using Step 3 and Step 4. This two step maximization is justified by Ma et al. (2015).

A.3.2. Pool-Adjacent-Violators Algorithm (PAVA)

We introduce a general concept of the pool-adjacent-violators algorithm (PAVA) before introducing the details of estimation method. Assume that $\mathbf{u} = (u_1, ..., u_n)$ are the observations and we want to find $\mathbf{v} = (v_1, ..., v_n)$ to minimize the following loss function $\sum_{i=1}^n (u_i - v_i)^2$ subject to $v_1 \leq v_2 \leq ... \leq v_n$. This regression is known as an isotonic regression. The isotonic regression can be easily solved by using a standard PAVA. Also, the PAVA can be interpreted as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$; $f_{\text{PAVA}} : \mathbb{R}^n \to \mathbb{R}^n$. If we have $\mathbf{u} = (u_1, ..., u_n)$, then we get the output $f_{\text{PAVA}}(\mathbf{u}) = \mathbf{v}$ as the solution for the minimization problem. More generally, if we have weights w_i, then the optimization problem is formulated by

$$\arg\min_{v_1\leq v_2\leq ...\leq v_n} \sum_{i=1}^n w_i(u_i - v_i)^2. \tag{A.3.1}$$

It is well-known that this optimization problem is easily solved by a weighted PAVA Barlow et al. (1972); de Leeuw et al. (2009).

A weighted PAVA is a simple algorithm that changes a sequence into an increasingly ordered sequence with predetermined weights. If we have a sequence $\mathbf{u} = (u_1, ..., u_n)$ and a weight sequence $\mathbf{w} = (w_1, ..., w_n)$, then the algorithm is as follows.

Step 1. Set $f_\uparrow(u_j) = u_j$ as initial values for all j.

Step 2. Start the first step with u_1. Take $f_\uparrow(u_1) = f_\uparrow(u_1)$ if $f_\uparrow(u_1) \leq f_\uparrow(u_2)$ and update $f_\uparrow(u_1) = f_\uparrow(u_2) = \frac{w_1 f_\uparrow(u_1) + w_2 f_\uparrow(u_2)}{w_1 + w_2}$ if $u_1 > u_2$. $\frac{w_1 f_\uparrow(u_1) + w_2 f_\uparrow(u_2)}{w_1 + w_2}$ is the weighted average of $\{f_\uparrow(u_1), f_\uparrow(u_2)\}$. Then, move to the next point $f_\uparrow(u_2)$. Note that the first

step does not update the points from the third to the last. That means $f_\uparrow(u_j) = u_j$ for $j = 3, ..., n$.

Step 3. For the i-th point, compare $f_\uparrow(u_i)$ with $f_\uparrow(u_{i+1})$. If $f_\uparrow(u_i) \leq f_\uparrow(u_{i+1})$, then $f_\uparrow(u_i)$ remains the same and move to the next point. If $f_\uparrow(u_i) > f_\uparrow(u_{i+1})$, then $f_\uparrow(u_i) = f_\uparrow(u_{i+1})$ is updated by the weighted average of $\{f_\uparrow(u_i), f_\uparrow(u_{i+1})\}$ and compare this value with $f_\uparrow(u_{i-1})$. If the nondecreasingness is achieved, i.e. $f_\uparrow(u_{i-1}) \leq f_\uparrow(u_i)$, then move to the $(i+1)$-th point. If $f_\uparrow(u_{i-1}) > f_\uparrow(u_i)$, then $f_\uparrow(u_{i-1}) = f_\uparrow(u_i) = f_\uparrow(u_{i+1})$ is updated by the weighted average of $\{f_\uparrow(u_{i-1}), f_\uparrow(u_i), f_\uparrow(u_{i+1})\}$. Keep doing this procedure until nondecreasingness is achieved. Once the partial sequence $(f_\uparrow(u_1), ..., f_\uparrow(u_i))$ is nondecreasing, then move to the $(i+1)$-th point.

The output $f_{\text{PAVA}}(\mathbf{u}) = (f_\uparrow(u_1), ..., f_\uparrow(u_n))$ from a weighted PAVA is nondecreasing and is the solution for the optimization problem (A.3.1). For a simple example, we consider a three-dimensional sequence of observations with equal weights. If we assume that the sequence is $\mathbf{u} = (3, 2, 1)$, then the output sequence $f_{\text{PAVA}}(\mathbf{u})$ is updated as the following order,

$$(3, 2, 1) \to (5/2, 5/2, 1) \to (5/2, 7/4, 7/4) \to (2, 2, 2).$$

In practice, there is a R package 'Iso' to implement a weighted PAVA. The 'pava' function with specifying weights can provide the output sequence $f_{\text{PAVA}}(\mathbf{u})$ for any sequence $\mathbf{u}$.

A.4. Appendix from Chapter 5

A.4.1. Evaluation of the choice of c_0

In Section 5.3, we discussed the penalized likelihood method using the penalty function $s(\alpha) = c_0\|\alpha\|$. Although our model is more complicated than the model in Efron (2016), the evaluation of the choice of c_0 is similar. To choose c_0, he considers the ratio of traces $R(\alpha)$, $R(\alpha) = \text{tr}\{\ddot{s}(\alpha)\}/\,\text{tr}\{\mathcal{I}(\alpha)\}$ where $\mathcal{I}(\alpha)$ is the Fisher information.

Like Efron (2016), assume that the space $\mathcal{T}$ of parasite density is a finite discrete set,

$$\mathcal{T} = \{z_1, \ldots, z_k\} \quad \text{with } z_1 = 0.$$

Then, the two densities $g_1(z)$ and $g_2(z)$ are re-written as

$$g_1(z_j; q, \alpha) = q \cdot I(z_j = 0) + (1-q) \cdot \exp\{Q_j^T\alpha - \phi_1(\alpha)\}I(z_j > 0), \quad 0 \leq q \leq 1$$
$$g_2(z_j; \alpha, \gamma) = \exp\{Q_j^T\alpha + \gamma z_j - \phi_2(\alpha, \gamma)\}$$

where Q_j^T is jth row of the known $k \times m$ structure matrix Q and α is a m-dimensional vector. We penalize the parameter α only, so we instead use $g_1(\alpha)$ and $g_2(\alpha)$ for simplicity. Also, since the measurement error mechanism h is known, then the vector $P_i = (p_{i1}, \ldots, p_{ik})^T$ for the observation D_i is also known where $p_{ij} = h(D_i; z_j)$. Furthermore, the marginal density

for D_i in the afebrile sample ($Y_i^{obs} = 0$) becomes

$$f_{0i}(D_i;\alpha) = \sum_{j=1}^{k} h(D_i; z_j) g_1(z_j;\alpha) = P_i^T g_1(\alpha).$$

Similarly, the marginal density in the febrile sample ($Y_i^{obs} = 1$) is

$$f_{1i}(D_i;\alpha) = (1-\lambda^*)(P_i^*)^T g_1(\alpha) + \lambda^* P_i^T g_2(\alpha).$$

where $P_i^* = (p_{i1}^*, \ldots, p_{ik}^*)^T$ and $p_{ij}^* = h(D_i; \beta z_j)$ with the fever killing parameter β. From Remark A2 in Efron (2016), we have

$$\begin{aligned}
\frac{\dot{f}_{0i}(\alpha)}{f_{0i}(\alpha)} &= Q^T W_{0i}(\alpha) \\
\frac{\dot{f}_{1i}(\alpha)}{f_{1i}(\alpha)} &= (1-\lambda^*) \cdot Q^T W_{0i}^*(\alpha) + \lambda^* \cdot Q^T W_{1i}(\alpha) \\
&= Q^T \left\{ (1-\lambda^*) W_{0i}^*(\alpha) + \lambda^* \cdot W_{1i}(\alpha) \right\} := Q^T \tilde{W}_{1i}(\alpha)
\end{aligned}$$

where

$$\begin{aligned}
W_{0i}(\alpha) &= g_1(z_j;\alpha) \left\{ p_{ij}/f_{0j}(\alpha) - 1 \right\} \\
W_{0i}^*(\alpha) &= g_1(z_j;\alpha) \left\{ p_{ij}^*/f_{1j}(\alpha) - 1 \right\} \\
W_{1i}(\alpha) &= g_2(z_j;\alpha) \left\{ p_{ij}/f_{1j}(\alpha) - 1 \right\}.
\end{aligned}$$

From Theorem 1 in Efron (2016), we can obtain the approximation for the ratio $R(\alpha)$ as $R(\hat{\alpha}) = \mathrm{tr}\{\ddot{s}(\hat{\alpha})\}/\,\mathrm{tr}\{\mathcal{I}(\hat{\alpha})\}$ where

$$\begin{aligned}
\ddot{s}(\hat{\alpha}) &= \frac{c_0}{\|\hat{\alpha}\|} \left(I - \frac{\hat{\alpha}\hat{\alpha}^T}{\|\hat{\alpha}\|^2} \right) \\
\mathcal{I}(\hat{\alpha}) &= \left[Q^T \left\{ \sum_{i=1}^{n_0} W_{0i}(\hat{\alpha}) \left\{ n_0 f_{0i}(\alpha) \right\} W_{0i}(\hat{\alpha})^T + \sum_{i=1}^{n_1} \tilde{W}_{1i}(\hat{\alpha}) \left\{ n_1 f_{1i}(\alpha) \right\} \tilde{W}_{1i}(\hat{\alpha})^T \right\} Q \right].
\end{aligned}$$

For the malaria example, $c_0 = 50$ was a modest choice for the regularizing constant since $R(\alpha) \approx 0.005$. For truncated normal distribution scenarios in simulations, $c_0 = 50$ was a modest choice.

A.4.2. Estimation of the dispersion parameter in the negative binomial distribution

In Section 6, we assume that the measurement error model (M2) has the negative binomial distribution: $D^{obs}|D^{cur} \sim 40 \times NB(D^{cur}/40, r)$ where the mean is $D^{cur}/40$ and the dispersion parameter is r. The dispersion parameter r is not known, but can be estimated from the data in O'Meara et al. (2007).

O'Meara et al. (2007) computed the false negative rate by counting numbers of slides reported as negative from 25 microscopists, and they plotted the false nagative rate on the

mean parasite density in Figure 2. We let y be the number of negative slides and x be the mean parasite density. We use the data (x, y) to find the maximum likelihood estimate of the dispersion parameter r; y_i is the number of 'negative' from the binomial distribution $B(n = 25, p_i)$ where p_i is the probability of being falsely negative, and p_i is computed from the negative binomial distribution $NB(x_i/40, r)$. The log-likelihood is given as

$$\begin{aligned} \ell &\propto \sum_{i=1}^{n} y_i \log(p_i) + (25 - y_i) \log(1 - p_i) \\ &= \sum_{i=1}^{n} y_i \log(f(0; x_i/40, r)) + (25 - y_i) \log(1 - f(0; x_i/40, r)) \end{aligned} \tag{A.4.1}$$

where $f(x; x_i/40, r)$ is the probability mass function of the negative binomial with the mean $x_i/40$ and the dispersion parameter r. From the data in O'Meara et al. (2007), the estimate of r is obtained as 5.83, and we use $r = 6$ in our paper. The R code for this estimation is provided online.

A.4.3. Connection to Existing Methods

We will show that the existing estimators of the MAFF are not consistent under Assumptions 2 and 3 alone, and that an additional, implausible assumption is needed. The estimator of the MAFF based on relative risk converges in probability to

$$\text{plim}(\widehat{MAFF}_{RR}) = p_f(R - 1)/R \tag{A.4.2}$$

where R is the relative risk of fever associated with the exposure of parasites, i.e. $R = P(Y^{obs} = 1|D^{obs} > 0)/P(Y^{obs} = 1|D^{obs} = 0)$. We note that p_a, p_f and R can be estimated from the observed data (Y^{obs}, D^{obs}). The consistency of the estimator $\widehat{MAFF}_{RR}$ relies on the following assumption.

Assumption 4. *No Errors Assumption.* The parasite density is not affected by a fever caused solely by a non-malaria infection, $Y^{nmi} = 1, Y^{mi} = 0$, and the observed parasite density D^{obs} is measured without error.

Proposition 1. Under Assumptions 2 - 4, the potential outcome framework MAFF is equal to the relative risk MAFF. That is, $MAFF_{potential} = \text{plim}(\widehat{MAFF}_{RR})$.

Proof. Since Y^{nmi} does not depend on the parasite level D and $P(Y^{mi} = 1|D = 0) = 0$, we can have

$$\begin{aligned} P(Y^{obs} = 1|D = 0) &= P(Y^{nmi} = 1, Y^{mi} = 0|D = 0) + P(Y^{mi} = 1|D = 0) \\ &= P(Y^{nmi} = 1)P(Y^{mi} = 0|D = 0) \\ &= P(Y^{nmi} = 1). \end{aligned} \tag{A.4.3}$$

Similarly, we have

$$\begin{aligned} P(Y^{obs} = 1|D > 0) &= P(Y^{nmi} = 1|D > 0) + P(Y^{nmi} = 0, Y^{mi} = 1|D > 0) \\ &= P(Y^{nmi} = 1) + P(Y^{nmi} = 0, Y^{mi} = 1|D > 0). \end{aligned} \tag{A.4.4}$$

Then, from Equations (A.4.3) and (A.4.4), $\widehat{MAFF}_{RR}$ is

$$\begin{aligned}
\widehat{MAFF}_{RR} &= p_f(R-1)/R \\
&= P(D>0|Y^{obs}=1)\cdot\frac{P(Y^{obs}=1|D>0)-P(Y^{obs}=1|D=0)}{P(Y^{obs}=1|D>0)} \\
&= P(D>0|Y^{obs}=1)\cdot\frac{\{P(Y^{nmi}=1)+P(Y^{nmi}=0,Y^{mi}=1|D>0)\}-P(Y^{nmi}=1)}{P(Y^{obs}=1|D>0)} \\
&= \frac{P(Y^{obs}=1,D>0)}{P(Y^{obs}=1)}\frac{P(Y^{nmi}=0,Y^{mi}=1|D>0)}{P(Y^{obs}=1|D>0)} \\
&= \frac{P(Y^{nmi}=0,Y^{mi}=1,D>0)}{P(Y^{obs}=1)} \\
&= P(Y^{nmi}=0,Y^{mi}=1|Y^{obs}=1).
\end{aligned}$$

□

Another popular choice of an estimator for the MAFF based on odds ratio is $\widehat{MAFF}_{OR} = (\hat{p}_f - \hat{p}_a)/(1-\hat{p}_a)$ that has the probability limit, $\text{plim}(\widehat{MAFF}_{OR})$. This estimator is an approximated version of the estimator $\hat{p}_f(\hat{R}-1)/\hat{R}$ because the relative risk R is often approximated by the odds ratio, $p_f(1-p_a)/p_a(1-p_f)$. The probability limit of this estimator is as

$$\text{plim}(\widehat{MAFF}_{OR}) = (p_f - p_a)/(1-p_a). \tag{A.4.5}$$

$\text{plim}(\widehat{MAFF}_{OR})$ is approximately equal to $\text{plim}(\widehat{MAFF}_{RR})$ when the prevalence of cases is rare.

Proposition 2. Under Assumptions 2-4, in terms of the potential outcome framework, $\text{plim}(\widehat{MAFF}_{OR})$ is given by

$$\text{plim}(\widehat{MAFF}_{OR}) = \frac{P(Y^{mi}=1)}{P(Y^{obs}=1)} = \frac{\text{plim}(\widehat{MAFF}_{RR})}{P(Y^{nmi}=0)}. \tag{A.4.6}$$

Proof. Let R^* be the odds ratio $p_f(1-p_a)/p_a(1-p_f)$. Since $p_f = P(D>0|Y^{obs}=1)$ and $p_a = P(D>0|Y^{obs}=0)$, the odds ratio R^* is

$$\begin{aligned}
R^* &= \frac{p_f}{1-p_f}\cdot\frac{1-p_a}{p_a} \\
&= \frac{P(D>0|Y^{obs}=1)}{P(D=0|Y^{obs}=1)}\cdot\frac{P(D=0|Y^{obs}=0)}{P(D>0|Y^{obs}=0)} \\
&= \frac{P(Y^{obs}=1,D>0)}{P(Y^{obs}=1,D=0)}\cdot\frac{P(Y^{obs}=0,D=0)}{P(Y^{obs}=0,D>0)} \\
&= \frac{P(Y^{obs}=1,D>0)}{P(Y^{obs}=0,D>0)}\cdot\frac{P(Y^{obs}=0,D=0)}{P(Y^{obs}=1,D=0)} \\
&= \frac{P(Y^{obs}=1|D>0)}{P(Y^{obs}=0|D>0)}\cdot\frac{P(Y^{obs}=0|D=0)}{P(Y^{obs}=1|D=0)}.
\end{aligned} \tag{A.4.7}$$

By substituting Equation (A.4.3), R^* is

$$\begin{aligned}
R^* &= \frac{P(Y^{obs}=1|D>0)}{P(Y^{nmi}=0,Y^{mi}=0|D>0)}\cdot\frac{P(Y^{nmi}=0)}{P(Y^{nmi}=1)} \\
&= \frac{P(Y^{obs}=1|D>0)}{P(Y^{nmi}=1,Y^{mi}=0|D>0)}.
\end{aligned} \tag{A.4.8}$$

Then, $\widehat{MAFF}_{OR}$ is

$$\begin{aligned}\widehat{MAFF}_{OR} &= p_f \cdot \tfrac{R^*-1}{R^*} \\ &= P(D>0|Y^{obs}=1) \cdot \tfrac{P(Y^{mi}=1|D>0)}{P(Y^{obs}=1|D>0)} \\ &= \tfrac{P(Y^{mi}=1,D>0)}{P(Y^{obs}=1)} \\ &= \tfrac{P(Y^{mi}=1)}{P(Y^{obs}=1)}\end{aligned} \tag{A.4.9}$$

□

According to Proposition 2, under Assumptions 2-4, $\widehat{MAFF}_{OR}$ is an asymptotically biased estimator of the MAFF, and estimates the proportion of children who have malaria fevers among febrile children $P(Y^{mi}=1|Y^{obs}=1)$, not the MAFF.

Proposition 2 implies that $\text{plim}(\widehat{MAFF}_{OR})$ is strictly larger than $\text{plim}(\widehat{MAFF}_{RR})$ when the probability of having non-malaria caused fever is positive. Technically, $\text{plim}(\widehat{MAFF}_{RR})$ is represented by multiplication of the probability of not having non-malaria caused fever $P(Y^{nmi}=0)$ and $\text{plim}(\widehat{MAFF}_{OR})$ as shown by equation (A.4.6). If the target estimand of some methods is $\text{plim}(\widehat{MAFF}_{OR})$, the estimate from the method should be adjusted by multiplying $P(Y^{nmi}=0)$ in order to acquire the estimate of $\text{plim}(\widehat{MAFF}_{RR})$. However, one difficulty is that $P(Y^{nmi}=0)$ is not observable. The following proposition shows that this adjustment can be successfully achieved by estimating $P(Y^{nmi}=0)$.

Proposition 3. Under Assumption 2-4, the estimator $\widehat{MAFF}_{RR}$ can be represented by the estimator $\widehat{MAFF}_{OR}$ and the probability $p = P(Y^{obs}=1)$. Let $\lambda = \widehat{MAFF}_{RR}$ and $\lambda^* = \widehat{MAFF}_{OR}$. Then, λ is obtained as

$$\lambda = \frac{\lambda^* - p\lambda^*}{1 - p\lambda^*}. \tag{A.4.10}$$

Proof. From equation (A.4.6),

$$\begin{aligned}\lambda &= P(Y^{nmi}=0) \cdot \lambda^* \\ &= \frac{P(Y^{mi}=0)P(Y^{nmi}=0)}{P(Y^{mi}=0)} \cdot \lambda^* \\ &= \frac{1-P(Y^{obs}=1)}{1-P(Y^{mi}=1)} \cdot \lambda^* \\ &= \frac{1-p}{1-p\lambda^*} \cdot \lambda^*.\end{aligned}$$

A.4.4. Simulation Study of Existing Methods

We evaluate the performance of several existing methods for estimating the MAFF with a simulation study. Specifically, we consider three settings; (1) there is no fever killing

Table 22: Means of estimates of the MAFF in Situation 1, 2 and 3 with 1000 simulations: Neither fever killing nor measurement error (Situation 1) No fever killing, but measurement error (Situation 2) and 50% fever killing and measurement error (Situation 3), True MAFF is 0.5.

	S	P	L	LI
Situation 1	0.499	0.449	0.470	0.507
Situation 2	0.469	0.442	0.386	0.476
Situation 3	0.334	0.286	0.258	0.370

and no measurement error (we call it Situation 1), (2) there is no fever killing effect, but there is measurement error (Situation 2) and (3) there is both fever killing (50%) and measurement error (Situation 3). We consider the four estimation methods discussed in Section 5.2: logistic regression (L), logistic regression with power parameter (P), local linear smoothing followed by isotonic regression (LI) and the adjusted semiparametric method (S). The first three methods (L, P and LI) have probability limits of plim($\widehat{MAFF}_{RR}$). The semiparametric model method has a probability limit of plim($\widehat{MAFF}_{OR}$), so we use adjustment (A.4.10) to obtain an estimate of plim($\widehat{MAFF}_{OR}$). We call this estimate the adjusted semiparametric method. The true MAFF is fixed as 0.5 across the simulation study; the true model is the first scenario described in Section 5.4 with sample size $n = 500$ and endemicity 0.2.

Table 22 shows the performance of the four estimators in the three situations. In Situation 1, the adjusted semiparametric method (S) and the nonparametric method (LI) produce estimates that are approximately unbiased; however, the other two methods produce biased estimates. This biased estimation for P and L is because the two methods rely on certain model assumptions and the true model in the simulation does not satisfy these model assumptions. In Situation 2 and 3, all estimators are significantly biased from the true value 0.5. In Situation 2, the increased magnitude of biases compared to Situation 1 can be understood as biases caused by measurement error. Also, the further increase in magnitude of bias in Situation 3 compared to Situation 2 represent biases caused by 50% fever killing. The combination of both fever killing and measurement error severely degrades the performance of the existing methods. Although the nonparametric method provides a good estimate of the MAFF in the absence of fever killing and measurement error, it performs poorly in the presence of both problems. The existing methods fail to provide unbiased estimates of the MAFF when Assumption 4 is violated.

Both the fever killing effect and measurement error are obstacles to obtain accurate measures of the parasitological challenge $D_i^{no.nmi}$ faced by a child. The failure to measure $D_i^{no.nmi}$ makes estimation of either $P(Y = 1|D^{no.nmi})$ or $f(D^{no.nmi}|Y = 1)$ biased, thus resulting in a biased estimate of the MAFF as can be seen in Table 22. In Section 5.4, more simulation results are displayed in various simulation settings. In the following section, we propose our new estimation method to account for both fever killing and measurement error by considering how to recover $D_i^{no.nmi}$ from D_i^{obs}.

BIBLIOGRAPHY

A. Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, 2002.

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

J. L. Aber, N. G. Bennett, D. C. Conley, and J. Li. The effects of poverty on child health and development. *Annual Review of Public Health*, 18(1):463–483, 1997.

L. H. Aiken, D. S. Havens, and D. M. Sloane. The magnet nursing services recognition program: A comparison of two groups of magnet hospitals. *The American Journal of Nursing*, 100(3):26–36, 2000. ISSN 0002-936X. URL http://journals.lww.com/ajnonline/Fulltext/2000/03000/The_Magnet_Nursing_Services_Recognition_Program__A.40.aspx.

J. D. Angrist and A. B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15 (4):69–85, 2001.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113. URL http://www.pnas.org/content/113/27/7353.abstract.

A. B. Atkinson. On the measuremnt of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

R. E. Barlow, D. J. Bartholomew, J. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions*. John Wiley & Sons, New York, 1972.

S. E. Barlow and W. H. Dietz. Obesity evaluation and treatment: expert committee recommendations. *Pediatrics*, 102(3):e29–e29, 1998.

B. Bergmann and G. Hommel. Improvements of general multiple test procedures for redundant systems of hypotheses. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 100–115. New York: Springer, 1988.

R. H. Berk and D. H. Jones. Goodness-of-fit test statistics that dominate the kolmogorov

statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

K. Bona, T. M. Blonquist, D. S. Neuberg, L. B. Silverman, and J. Wolfe. Impact of socioeconomic status on timing of relapse and overall survival for children treated on dana-farber cancer institute all consortium protocols (2000–2010). *Pediatric Blood & Cancer*, 63(6):1012–1018, 2016.

C. S. Boutlis, T. W. Yeo, and N. M. Anstey. Malaria tolerance–for whom the cell tolls? *Trends in Parasitology*, 22(8):371–377, 2006.

P. Bouvier, A. Rougemont, N. Breslow, O. Doumbo, V. Delley, A. Dicko, M. Diakite, A. Mauris, and C.-F. Robert. Seasonality and malaria in a west african village: does high parasite density predict fever incidence? *American Journal of Epidemiology*, 145 (9):850–857, 1997.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

M. A. Brookhart and S. Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, 3(1):Article 14, 2007.

J. Cheng, J. Qin, and B. Zhang. Semiparametric estimation and inference for distributional and genenral treatment effects. *Journal of the Royal Statistical Society: Series B*, 71: 881–904, 2009.

A. Chesher. Testing for neglected heterogeneity. *Econometrica*, 52(4):865–872, 1984.

Coalition PM. The case for personalized medicine. Technical report, Coalition for personalized medicine, 2014. URL http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/pmc_the_case_for_personalized_medicine.pdf.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of National Cancer Institute*, 22(1):173–203, 1959.

D. R. Cox. *Analysis of Binary Data*. London, Methuen, 1970.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, 90(3):389–405, Aug. 2008. ISSN 0034-6535. doi: 10.1162/rest.90.3.389. URL http://dx.doi.org/10.1162/rest.90.3.389.

M. A. Davis, J. M. Neuhaus, D. J. Moritz, D. Lein, J. D. Barclay, and S. P. Murphy. Health behaviors and survival among middle aged and older men and women in the nhanes i epidemiologic follow-up study. *Preventive medicine*, 23(3):369–376, 1994.

J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B*, 78(3):655–671, 2015.

M. Dowling and G. Shute. A comparative study of thick and thin blood films in the diagnosis of scanty malaria parasitaemia. *Bulletin of the World health Organization*, 34 (2):249, 1966.

W. C. Earle, M. Perez, et al. Enumeration of parasites in the blood of malarial patients. *Journal of Laboratory and Clinical Medicine*, 17(11):1124–1130, 1932.

B. Efron. Empirical Bayes deconvolution estimates. *Biometrika*, 103(1):1–20, 2016. URL http://biomet.oxfordjournals.org/content/103/1/1.abstract.

B. L. Egleston, D. O. Scharfstein, and E. MacKenzie. On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, 65(2):497–504, 2009.

R. A. Fisher. *The Design of Experiments*. Edinburgh: Oliver & Boyd, 1935.

C. B. Fogarty and D. S. Small. Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association*, 111(516):1820–1830, 2016.

J. L. Gastwirth. Methods for assessing the sensitivity of statistical comparisons used in title vii cases to omitted variables. *Jurimetrics*, 33(1):19–34, 1992.

J. L. Gastwirth, A. M. Krieger, and P. R. Rosenbaum. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 62 (3):545–555, 2000.

S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*, volume 195. New York: Springer, 2009.

P. B. Gilbert, R. J. Bosch, and M. G. Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541, 2003.

B. M. Greenwood, A. Bradley, A. Greenwood, P. Byass, K. Jammeh, K. Marsh, S. Tulloch, F. Oldfield, and R. Hayes. Mortality and morbidity from malaria among children in a rural area of the gambia, west africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 81(3):478–486, 1987.

M. E. Halloran, I. M. Longini Jr, and C. J. Struchiner. Design and interpretation of vaccine field studies. *Epidemiologic Reviews*, 21(1):73–88, 1999.

M. A. Hamburg and F. S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

D. M. Haughton et al. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

P. J. Heagerty and S. R. Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.

R. Heller, P. R. Rosenbaum, and D. S. Small. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101, 2009.

M. A. Hernan and J. M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

P. W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18:449–484, 1988.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

M. Hommel. Diagnostic methods in malaria. *Essential malariology. 4th ed. London: Edward Arnold*, pages 35–56, 2002.

C. A. Hosman, B. B. Hansen, and P. W. Holland. The sensitivity of linear regression coefficients'confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849–870, 2010.

J. Y. Hsu, D. S. Small, and P. R. Rosenbaum. Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108(501):135–148, 2013.

J. Y. Hsu, J. R. Zubizarreta, D. S. Small, and P. R. Rosenbaum. Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102(4):767–782, 2015.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994a.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994b.

G. W. Imbens and D. B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

J. Jacobson, R. Briefel, P. Gleason, and R. Sullivan. Designs for measuring how the school breakfast program affects learning. *Mathematica Policy Research, Inc*, 2001.

K. Jogdeo. Association and probability inequalities. *The Annals of Statistics*, 5(3):495–504, 1977.

E. Karni. A theory of medical decision making under uncertainty. *Journal of Risk and Uncertainty*, 39(1):1–16, 2009.

S. Kitchen. Falciparum malaria. *Malariology. Philadelphia: WB Saunders*, 2:995–1016, 1949.

K. A. Koram and M. E. Molyneux. When is "malaria" malaria? the different burdens of malaria infection, malaria disease, and malaria-like illnesses. *The American Journal of Tropical Medicine and Hygiene*, 77(6 Suppl):1–5, 2007.

D. Kwiatkowski. Febrile temperatures can synchronize the growth of plasmodium falciparum in vitro. *The Journal of Experimental Medicine*, 169(1):357–361, 1989.

D. Kwiatkowski and M. Nowak. Periodic and chaotic host-parasite interactions in human malaria. *Proceedings of the National Academy of Sciences*, 88(12):5111–5113, 1991.

F. Larribe and P. Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43, 2011.

E. L. Lehmann. *Nonparametrics*. San Francisco: Holden-Day, 1975.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. New York: Springer, 2005.

S. F. Lehrer, R. V. Pohl, and K. Song. Targeting policies: Multiple testing and distributional treatment eects. Working Paper 22950, National Bureau of Economic Research, December 2016. URL http://www.nber.org/papers/w22950.

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39, 1988.

W. Liu, S. J. Kuramoto, and E. A. Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14(6):570–580, 2013. ISSN 1573-6695. URL http://dx.doi.org/10.1007/s11121-012-0339-5.

H. Long, B. Lell, K. Dietz, and P. Kremsner. Plasmodium falciparum: in vitro growth inhibition by febrile temperatures. *Parasitology research*, 87(7):553–555, 2001.

X. Lu and H. White. Testing for treatment dependence of effects of a continuous treatment. *Econometric Theory*, 31(5):1016–1053, 2015.

Z. Ma, D. Foster, and R. Stine. Adaptive monotone shrinkage for regression. *arXiv preprint arXiv:1505.01743*, 2015.

S. Mabunda, J. J. Aponte, A. Tiago, and P. Alonso. A country-wide malaria survey in mozambique. ii. malaria attributable proportion of fever and establishment of malaria case definition in children across different epidemiological settings. *Malaria Journal*, 8 (1):74, 2009. ISSN 1475-2875. doi: 10.1186/1475-2875-8-74. URL http://dx.doi.org/10.1186/1475-2875-8-74.

R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

J. Maritz. A note on exact robust confidence intervals for location. *Biometrika*, 66(1): 163–166, 1979.

K. Marsh. Immunology of malaria. *London: Arnold Publishers*, pages 252–265, 2002.

F. E. McKenzie, W. A. Prudhomme, A. J. Magill, J. R. Forney, B. Permpanich, C. Lucas, R. A. Gasser, and C. Wongsrichanalai. White blood cell counts and malaria. *Journal of Infectious Diseases*, 192(2):323–330, 2005.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1923, 1990.

P. C. O'Brien and T. R. Fleming. A paired prentice-wilcoxon test for censored paired data. *Biometrics*, 43(1):169–180, 1987.

E. L. Ogburn, A. Rotnitzky, and J. M. Robins. Doubly robust estmation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B*, 77: 373–396, 2015.

C. L. Ogden and K. M. Flegal. Changes in terminology for childhood overweight and obesity. *National Health Statistics Reports*, 25:1–5, 2010.

W. P. O'Meara, B. F. Hall, and F. E. McKenzie. Malaria vaccine efficacy: the difficulty of detecting and diagnosing malaria. *Malaria Journal*, 6(1):1–10, 2007.

A. B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC Press, 2001.

J. Qin and D. H. Leung. A semiparametric two-component "compound" mixture model and its application to estimating malaria attributable fractions. *Biometrics*, 61(2):456–464, 2005.

K. Ralston, C. Newman, A. Clauson, J. Guthrie, and J. Buzby. The national school lunch

program: Background, trends, and issues. economic research report number 61. *US Department of Agriculture*, 2008.

C. Rogier, D. Commenges, and J.-F. Trape. Evidence for an age-dependent pyrogenic threshold of plasmodium falciparum parasitemia in highly endemic populations. *The American Journal of Tropical Medicine and Hygiene*, 54(6):613–619, 1996.

I. B. Rooth and A. Bjorkman. Suppression of plasmodium falciparum infections during concomitant measles or influenza but not during pertussis. *The American Journal of Tropical Medicine and Hygiene*, 47(5):675–681, 1992.

P. R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B*, 53(3):597–610, 1991.

P. R. Rosenbaum. Attributing effects to treatment in matched observational studies. *Journal of the American statistical Association*, 97(457):183–192, 2002a.

P. R. Rosenbaum. *Observational Studies*. New York: Springer, 2002b.

P. R. Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

P. R. Rosenbaum. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152, 2005.

P. R. Rosenbaum. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464, 2007.

P. R. Rosenbaum. *Design of Observational Studies*. New York: Springer, 2010.

P. R. Rosenbaum. Testing one hypothesis twice in observational studies. *Biometrika*, 99 (4):763–774, 2012.

P. R. Rosenbaum. Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, 110(509):205–217, 2015.

P. R. Rosenbaum and J. H. Silber. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405, 2009a.

P. R. Rosenbaum and J. H. Silber. Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *Journal of the American Statistical Association*, 104(486):501–511, 2009b.

P. R. Rosenbaum and D. S. Small. An adaptive mantel–haenszel test for sensitivity analysis in observational studies. *Biometrics*, 2016.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

J. P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831, 1986.

J. H. Silber, P. R. Rosenbaum, M. D. McHugh, and et al. Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surgery*, 151(6):527–536, 2016. ISSN 2168-6254. doi: 10.1001/jamasurg.2015.4908. URL http://dx.doi.org/10.1001/jamasurg.2015.4908.

D. S. Small, J. Cheng, and T. R. Ten Have. Evaluating the efficacy of a malaria vaccine. *The International Journal of Biostatistics*, 6(2):1–22, 2010.

T. Smith, J. A. Schellenberg, and R. Hayes. Attributable fraction estimates and case definitions for malaria in endemic. *Statistics in Medicine*, 13(22):2345–2358, 1994.

T. A. Smith. Measures of clinical malaria in field trials of interventions against plasmodium falciparum. *Malaria Journal*, 6(1):53, 2007.

M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *journal of Business and Economic Statistics*, 20(4):518–529, 2002.

M. Story, K. M. Kaphingst, R. Robinson-O'Brien, and K. Glanz. Creating healthy food and eating environments: policy and environmental approaches. *Annual Review of Public Health*, 29:253–272, 2008.

E. A. Stuart and D. B. Hanna. Commentary: Should epidemiologists be more sensitive to design sensitivity? *Epidemiology*, 24(1):88–89, 2013.

Z. Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

M. J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23, 2004.

A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer, 1996.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

P. Vounatsou, T. Smith, and A. Smith. Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):575–587, 1998.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2015.

S. Walter. The estimation and interpretation of attributable risk in health research. *Biometrics*, 32(4):829–849, 1976.

W. Wang and D. Small. A comparative study of parametric and nonparametric estimates of the attributable fraction for a semi-continuous exposure. *The International Journal of Biostatistics*, 8(1), 2012.

D. Warrell. *Clinical features of malaria*. London: Arnold, 1993.

N. Wertheimer and E. Leeper. Electrical wiring configurations and childhood cancer. *American Journal of Epidemiology*, 109(3):273–284, 1979.

D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.

H. Zhang and B. Singer. *Recursive partitioning and applications*. New York: Springer, 2010.

J. R. Zubizarreta, M. Neuman, J. H. Silber, and P. R. Rosenbaum. Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association*, 107(499): 901–915, 2012.

J. R. Zubizarreta, M. Cerdá, and P. R. Rosenbaum. Effect of the 2010 chilean earthquake on posttraumatic stress reducing sensitivity to unmeasured bias through study design. *Epidemiology*, 24(1):79–87, 2013.