STATISTICAL METHODS FOR STUDYING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES

by

Michael Johnson

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 12/6/21

The dissertation is approved by the following members of the Final Oral Committee:
Hyunseung Kang, Assistant Professor, Department of Statistics
Menggang Yu, Professor, Department of Biostatistics and Medical Informatics
Guanhua Chen, Assistant Professor, Department of Biostatistics and Medical Informatics
Jun Zhu, Professor, Department of Statistics
Lu Mao, Assistant Professor, Department of Biostatistics and Medical Informatics

To my brother Matthew, in loving memory.

ACKNOWLEDGMENTS

It has been a challenging but ultimately rewarding journey toward a PhD in Statistics, and would not have been possible without the considerable help and support I received along the way (I don't need to see the counterfactual to know that's causal).

First and foremost, I would like to extend my deepest sense of gratitude to my advisor, Professor Hyunseung Kang, and my co-advisors, Professors Menggang Yu and Guanhua Chen, for their patience, guidance, support, understanding, and encouragement. It was Professor Kang who introduced me to causal inference, instrumental variables, and heterogeneous treatment effects, which resulted in opportunities to travel the world and present my work in three different countries. His calm and cool demeanor made it easy to ask questions and discuss ideas, and his sharp insight provided much inspiration and refined my proposed innovations. Professors Yu and Chen introduced me to personalized medicine and gave me a deeper appreciation for independent research, both in what it entails and how to conduct it. I am eternally grateful for each of my advisors' mentorship in matters academic, professional, and personal.

To Professors Kang, Yu, and Chen, without your feedback, suggestions, and guidance, my dissertation wouldn't have been possible. I admire you all as both mentors in profession and as role models in life.

I would like to also thank my dissertation committee members, Professors Jun Zhu and Lu Mao for their constructive comments and suggestions. I recall the sincere support I received from Professor Zhu at a time I was questioning my place in the Statistics PhD program; I may not be where I am if it weren't for that help.

I would like to thank my friends both in and out of Madison. They made the journey all the more memorable and helped to provide some much needed respite. Though I am thankful to all of my friends that I've made along the way, in particular, I would like to thank: Jacob Maronge and John Spaw for being with me from the beginning (good morning), Behzad Aalipur, Fan Chen, Lisa Gao, Liam Johnston, and Victor Luo for time spent studying and relaxing together, and Jungjae Park, Benjamin Fincher, Rui Shi, Shabab Siddiqui, Matthew Dotray, Jeremy Tow, Tanli Sun, and Jack Shen for always being there for me and at least feigning interest in my research.

I am also eternally grateful for the encouragement and support of my loving family. I would like to thank each of my siblings Kristen, Katharine, and Mark for their support

and care. Finally, I would like to especially thank my partner Lan Luo, for her unwavering support, comfort, patience, and encouragement, and my parents Mark and Patty, for their unconditional love, continued confidence in me, and devoting so much time, and effort to me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

There is a growing interest in estimating heterogeneous treatment effects in randomized and observational studies. However, most of the work relies on the assumption of ignorability, or no unmeasured confounding on the treatment effect. While instrumental variables (IV) are a popular technique to control for unmeasured confounding, there has been little research conducted to study heterogeneous treatment effects with the use of an IV. This dissertation introduces methods using an IV to discover novel subgroups, estimate their heterogeneous treatment effects, and identify individualized treatment rules (ITR) when ignorability is expected to be violated.

In Chapter 2, we present a two-part algorithm to estimate heterogeneous treatment effects and detect novel subgroups using an IV with matching. The first part uses interpretable machine learning techniques, such as classification and regression trees, to discover potential effect modifiers. The second part uses closed testing to test for statistical significance of each effect modifier while strongly controlling the familywise error rate. We apply this method on the Oregon Health Insurance Experiment, estimating the effect of Medicaid on the number of days an individual's health does not impede their usual activities by using a randomized lottery as an instrument.

In Chapter 3, we generalize methods to identify ITR using a binary IV to using multiple, discrete valued instruments, or equivalently, multilevel instruments. Several new problems arise when generalizing to multilevel instruments, requiring novel solutions. In particular, multilevel IV give rise to many latent subgroups that may experience heterogeneous treatment effects. Additionally, it may be unclear how to combine and compare the different levels of the IV to estimate treatment heterogeneity. We provide methods that use a prediction of the latent subgroup to identify optimal ITR, and methods to dynamically combine levels of the multilevel IV to estimate the heterogeneous treatment effects, effectively individualizing estimation of an ITR. Further, we provide and discuss necessary and sufficient conditions to identify an optimal ITR using a multilevel IV. We apply our methods to identify an ITR for two competing treatments, carotid endarterectomy and carotid artery stenting, on preventing stroke or death within 30 days of their index procedure.

1 INTRODUCTION

When comparing the effectiveness of two or more treatments, in addition to assessing the overall treatment effects, it is often of interest to analyze treatment effects within subgroups. Traditionally, approaches to study the heterogeneous treatment effects (i.e. the various treatment effects exhibited by the subgroups) required prespecification of subpopulations rather than using data to suggest unknown subgroups [Stallones (1987), Yusuf et al. (1991), CPMP (1995), Rothwell (2005)]. However, particularly in the era of Big Data [Hilbert and López (2011)], there exists too many possible subgroups for domain experts to prespecify, necessitating the development of methods that can detect novel subgroups and estimate the heterogeneous treatment effects from data. Recently, there have been extensive research on techniques to identify data-driven subgroups and estimate effect heterogeneity; see Su et al. (2009), Hill (2011), Hsu et al. (2013), Tian et al. (2014), Hsu et al. (2015), Athey and Imbens (2016), Lee et al. (2018b), Lee et al. (2018a), Wager and Athey (2018), Chernozhukov et al. (2018), Hahn, Murray, and Carvalho (2020), Wang and Rudin (2021), Lee et al. (2021a) and references therein. Further, as many practitioners and policy-makers are most interested in using the understood subgroups and estimated heterogeneous effects to inform decisions, the subgroup-specific treatment effects can be used to develop decision strategies to provide the treatment most appropriate for the subgroup. Such decision strategies are referred to as optimal decision rules, or individualized treatment rules (ITR). Recently, there has been much progress made in developing methods identifying such rules; see Zhao et al. (2009), Qian and Murphy (2011), Zhao et al. (2012), Zhang et al. (2012), Moodie et al. (2014), Xu et al. (2015), Chen et al. (2016), Chen et al. (2017), Sutton and Barto (2018), Lou et al. (2018), Chen et al. (2018) and references therein. However, the aforementioned works rely on the assumption of ignorability, that there is no unmeasured confounding on the effect of the treatment on the outcome. Unfortunately, this condition may not hold in observational studies or randomized trials with noncompliance.

Instrumental variables (IV) are variables related to the outcome only through the treatment and methods using an IV were developed to control for unmeasured confounding. The general idea of an IV method has two parts, (i) find a variable, called an instrument, that influences the treatment of interest, but is otherwise not associated with the outcome, and is independent of unmeasured confounders, and (ii) use the variation free of unmeasured confounding in the treatment, elicited by the IV, to estimate causal effects of

the treatment. Some commonly used IVs are assignment to a treatment in randomized trials with noncompliance, randomized encouragement to take a treatment, randomized lotteries, regression discontinuity designs, distance to specialty care providers, physician or hospital preference, and multiple genetic variants; for additional discussions and review, see Angrist et al. (1996), Hernán and Robins (2006), and Baiocchi et al. (2014). Estimation and inference of causal effects using IV have been extensively studied, mostly using likelihood, series, sieve, minimum distance, matching, and/or moment-based methods; see Abadie (2003); Blundell and Powell (2003); Newey and Powell (2003); Ai and Chen (2003); Hall and Horowitz (2005); Tan (2006); Blundell et al. (2007); Baiocchi et al. (2010); Kang et al. (2016a); Darolles et al. (2011); Chen and Pouzo (2012); Okui et al. (2012); Su et al. (2013); Athey et al. (2019) and references therein. However, work on developing methods for estimating and inferring heterogeneous treatment effects using IV is understudied.

Recently, there has been some work to incorporate IV techniques to identify optimal ITR. Methods estimating heterogeneous treatment effects and identifying subgroups using IV and tree-based methods have been proposed by Bargagli-Stoffi and Gnecco (2018), Bargagli-Stoffi et al. (2019), and Athey et al. (2019). An extension of Outcome Weighted Learning, a popular technique for identifying ITR, to include IV was provided by Cui and Tchetgen Tchetgen (2021b). Alternative approaches estimating heterogeneous treatment effects with an IV to identify an ITR were presented by Pu and Zhang (2020) where different assumptions surrounding the instrument were considered, and Qiu et al. (2021) where settings with restrictions placed on resources were examined. Time varying treatment rules, or dynamic treatment regimes, using an IV were proposed by Chen and Zhang (2021). For further discussions on identifying ITR with IV, see Cui and Tchetgen Tchetgen (2021a). However, no techniques have used matching, a popular and intuitive method in causal inference, to nonparametrically estimate treatment heterogeneity and identify novel subgroups with the use of an IV. Further, all of the methods identifying ITR consider the use of a binary IV, which, as many of the IVs commonly used in practice are not binary, can be a limitation in practice.

The focus of this dissertation is to provide additional techniques to estimate heterogeneous treatment effects, identify novel subgroups, and develop optimal ITR using IV and provide practical guidance on the use of such methods. Specifically, in Chapter 2, an algorithm to discover novel subgroups and estimate heterogeneous treatment effects using an IV is proposed. In this algorithm, we incorporate IV with matching, classification and regression trees, and closed testing to provide inferential guarantees by strongly controlling

familywise error rates. In numerical examples, we demonstrate the algorithm's capability to detect existing subgroups while maintaining honest inference (i.e. controlling the familywise error rate). Finally, We apply our method to the Oregon Health Insurance Experiment, a natural experiment, to study heterogeneous treatment effects of Medicaid among the complier subpopulation, while adding details on how to perform similar analyses. In Chapter 3, we introduce methods to identify optimal ITR in the presence of unmeasured confounding using multiple, discrete valued instruments, or equivalently, multilevel instruments. In particular, we provide extensions of the existing literature using binary IV and discussion on the assumptions required to identify ITR using an IV. As extending to multilevel IV introduces several aspects not considered when using a binary IV, we also provide novel methods that use observed and latent effect modifiers and methods that individualize estimation of ITR. We additionally provide extensive discussion on how to use the methods, specifically in the decisions on how to use the different levels of the IV. In simulation studies, we demonstrate how using a multilevel IV over its natural support improves estimation over the use of a dichotomized IV. Lastly, we apply our methods to data from the Vascular Quality Initiative (`https://www.vqi.org`) to develop an optimal ITR between competing treatments for patients at risk of stroke or death from carotid artery disease. The techniques detailed in this dissertation help to fill the methodology gaps in studying heterogeneous treatment effects with an IV.

2 DETECTING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES AND APPLICATION TO THE OREGON HEALTH INSURANCE EXPERIMENT

2.1 Introduction

2.1.1 Motivation: Utilization of Medicaid in Oregon and the Complier Average Causal Effect

In January of 2008, Oregon reopened its Medicaid-based health insurance plan for its eligible residents and, for a brief period, allowed a limited number of individuals to enroll in the program. Specifically, a household in Oregon was randomly selected by a lottery system run by the state and any eligible individual in the household can choose to enroll in the new health insurance plan; households that weren't selected by the lottery could not enroll whatsoever.

For policymakers, Oregon's randomized lottery system was a unique opportunity, specifically a natural experiment, to study Medicaid's causal effect on a variety of health and economic outcomes, as directly randomizing Medicaid (or withholding it) to individuals would be infeasible and unethical. In this natural experiment, commonly referred to as the Oregon Health Insurance Experiment (OHIE), Finkelstein et al. (2012) used the randomized lottery as an instrumental variable (see Section 2.2.2 for details) to study the complier average causal effect (CACE), or the effect of Medicaid among individuals who enrolled in Medicaid after winning the lottery (Angrist, Imbens, and Rubin, 1996). The CACE reflects Medicaid's impact among a subgroup of individuals and differs from the average treatment effect for the entire population (ATE) or the intent-to-treat (ITT) effect of the lottery itself on the outcome. In this paper, we focus on studying the CACE; see Imbens (2010), Swanson and Hernán (2013), and Swanson and Hernán (2014) for additional discussions on the CACE.

Often in studying the CACE, the population of compliers is assumed to be homogeneous whereby two compliers are alike and have the same treatment effect. But, no two individuals are the same and it is plausible that some compliers may benefit more from the treatment than other compliers. For example, sick individuals who enroll in Medicaid after winning the lottery may benefit more from Medicaid than healthy individuals. Also, the perceived benefit

of enrolling in Medicaid among sick versus healthy individuals may create heterogeneity in the compliance rate, i.e. the number of people who sign up when they win the lottery, with sick people presumably signing up more than healthy people. Alternatively, if people are equally likely to enroll in Medicaid when they win the lottery, those who are unemployed may benefit more from Medicaid in terms of reducing out-of-pocket healthcare spending and medical debt than those who are employed. The theme of this paper is to explore these issues, specifically the heterogeneity of CACE and how to discover them in an honest manner by using well-known matching methods and recent tree-based methods in heterogeneous treatment effect estimation.

2.1.2 Prior Work and Our Contributions

Traditional approaches to study heterogeneous effects required subgroups to be specified a priori rather than allowing for unknown subgroups to be discovered by the data (Stallones, 1987; Yusuf et al., 1991; Rothwell, 2005). In recent years, there have been many works in causal inference using tree-based methods to estimate effect heterogeneity or to identify data-driven subgroups when there is full compliance; see Su et al. (2009), Hill (2011), Athey and Imbens (2016), Wager and Athey (2018), Chernozhukov et al. (2018), Athey, Tibshirani, and Wager (2019), Hahn, Murray, and Carvalho (2020), Wang and Rudin (2021), Lee et al. (2021a) and references therein. Notably, Wang and Rudin (2021), Lee et al. (2021b), and Lee et al. (2021a) used data to suggest novel effect modifiers, aiding domain experts identify new subgroups when there are too many possible subgroups to consider. The majority of the aforementioned work utilizes sample splitting or sub-sampling to obtain honest inference. Here, honest inference refers to a procedure that controls the Type I error rate (or the familywise error rate) of testing a null hypothesis about a treatment effect at a desired level α; see Section 2.2.4 for additional discussions. However, Hsu, Small, and Rosenbaum (2013) used pair matching and classification and regression trees (CART) (Breiman et al., 1984) to conduct honest inference, all without sample splitting. A follow-up work by Hsu et al. (2015) formally showed that the procedure strongly controls the familywise error rate for testing heterogeneous treatment effects, again without sample splitting. Subsequent works by Lee et al. (2018a), Lee et al. (2018b), and Lee, Small, and Dominici (2021b) extended this idea to increase statistical power of detecting such effects.

There is also work on nonparametrically estimating treatment effects using instrumental variables (IV), mostly using likelihood, series, sieve, minimum distance, and/or moment-

based methods; see Abadie (2003); Blundell and Powell (2003); Newey and Powell (2003); Ai and Chen (2003); Hall and Horowitz (2005); Blundell et al. (2007); Darolles et al. (2011); Chen and Pouzo (2012); Su et al. (2013); Athey et al. (2019) and references therein. Recently, Bargagli-Stoffi and Gnecco (2018) and Bargagli-Stoffi et al. (2019) explored effect heterogeneity in the CACE by using causal trees (Athey and Imbens, 2015) and Bayesian causal forests (Hahn et al., 2020), specifically by estimating heterogeneity in the ITT effect and dividing it by the compliance rate. However, to the best of our knowledge, none have used matching, a popular, intuitive, and easy-to-understand method in causal inference, as a device to nonparametrically estimate treatment heterogeneity in the CACE and to guarantee strong familywise Type I error control. Works on using matching with an instrument by Baiocchi et al. (2010) and Kang et al. (2013, 2016a) only focused on the population CACE; they do not explore heterogeneity in the CACE. Also, aforementioned works by Hsu et al. (2013) and Hsu et al. (2015) using matching and CART did not consider instruments.

The goal of this paper is to propose a matching-based method to study effect heterogeneity and identify novel, data-driven subgroups in instrumental variables settings. Specifically, the target estimand of interest is what we call the *heterogeneous* complier average causal effect (H-CACE). A heterogeneous complier average causal effect (H-CACE) is the usual complier average causal effect, but for a subgroup of individuals defined by their pre-instrument covariates. At a high level, H-CACE explores treatment heterogeneity in the complier population, where we suspect that not all compliers in the data react to the treatment in the same way. Some subgroup of compliers may respond to the treatment differently than another subgroup of compliers, who may not respond to the treatment at all; some may even be more likely to be compliers if they believe the treatment would benefit them and they may actually benefit from the treatment. The usual CACE obscures the underlying heterogeneity among compliers by averaging across different types of compliers whereas H-CACE attempts to expose it. Also, in the case where the four compliance types in Angrist et al. (1996), specifically compliers, never-takers, always-takers, and defiers, have identical effects, the H-CACE can identify the heterogeneous treatment effect for the entire population using an instrument. Section 2.2.3 formalizes H-CACE and provides additional discussions.

Methodologically, to study H-CACE, we combine existing ideas of heterogeneous treatment effect estimation in non-IV matching contexts by Hsu et al. (2015) and matching with IVs by Baiocchi et al. (2010) and Kang et al. (2016a). Specifically, we first follow Baiocchi et al. (2010) and Kang et al. (2016a) and conduct pair matching on a set of pre-

instrument covariates. Second, we follow Hsu et al. (2015) where we obscure the difference in the outcomes between treated and controls by using absolute differences and use CART to discover novel subgroups of study units without contaminating downstream inference. Specifically, we use closed testing to test the H-CACE in different subgroups while strongly controlling for familywise error rate (Marcus, Eric, and Gabriel, 1976). Simulation studies are conducted to evaluate the performance of our proposed method under varying levels of compliance and effect heterogeneity. The simulation study also compares our method to the recent aforementioned method by Bargagli-Stoffi et al. (2019). We then use our method to analyze heterogeneity in the effect of Medicaid on increasing the number of days a complying individual's health does not hamper their usual activities.

2.2 Methodology

2.2.1 Notation

Let $i = 1, \ldots, I$ index the I matched pairs and $j = 1, 2$ index the units within each matched pair i. Let Z_{ij} be a binary instrument for unit j in matched pair i where one unit in the pair receives the instrument value $Z_{ij} = 1$ and the other receives the value $Z_{ij} = 0$. In the OHIE data, $Z_{ij} = 1$ and $Z_{ij} = 0$ denotes an individual winning or losing the Medicaid lottery, respectively. Let $\mathbf{Z}$ be the vector of instruments, $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I1}, Z_{I2})$ and $\mathcal{Z}$ denote an event of instrument assignments for all units.

For unit j in matched pair i, let d_{1ij} and d_{0ij} denote the binary potential treatment/exposure given the instrument value of $Z_{ij} = 1$ and $Z_{ij} = 0$ respectively. Further, define the potential response $r_{1ij}^{(d_{1ij})}$ for unit j in matched set i with exposure d_{1ij} receiving instrument value $Z_{ij} = 1$; we define $r_{0ij}^{(d_{0ij})}$ similarly but with instrument value $Z_{ij} = 0$. For the OHIE data, d_{1ij} denotes whether an individual enrolled in Medicaid and $r_{1ij}^{(d_{1ij})}$ denotes the potential outcome when the individual wins the lottery $Z_{ij} = 1$. For unit j in matched set i, the observed response is defined as $R_{ij} = r_{1ij}^{(d_{1ij})} Z_{ij} + r_{0ij}^{(d_{0ij})}(1 - Z_{ij})$ and the observed treatment is defined as $D_{ij} = d_{1ij} Z_{ij} + d_{0ij}(1 - Z_{ij})$. The notation assumes that the Stable Unit Treatment Value Assumption (SUTVA) holds (Rubin, 1980). Define $\mathcal{F} = \{(r_{1ij}^{(d_{1ij})}, r_{0ij}^{(d_{0ij})}, d_{1ij}, d_{0ij}, \mathbf{X}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$ to be the set of potential outcomes, treatments, and covariates, both observed, $\mathbf{X}_{ij}$, and unobserved, u_{ij}.

When partitioning the matched sets into subgroups for discovering effect heterogeneity, the following notation is used. We define a "set of sets", or grouping $\mathcal{G}$, which contains mu-

tually exclusive and exhaustive subsets of the pairs $s_g \subseteq \{1, \ldots, I\}$ so that $\mathcal{G} = \{s_1, \ldots, s_G\}$. The subscript g in s_g is used to denote a unit partitioned into the gth subset s_g. To avoid overloading the notation, s and s_g will be used interchangeably when it isn't necessary to specify a subgroup g. The set of potential outcomes, treatments, and covariates for subset s_g are defined as $\mathcal{F}_{s_g} = \{(r_{1sij}^{(d_{1sij})}, r_{0sij}^{(d_{0sij})}, d_{1sij}, d_{0sij}, \mathbf{X}_{sij}, u_{sij}) : s_g \subseteq \{1, \ldots, I\}, i \in s_g, j = 1, 2\}$, where $\mathcal{F} = \bigcup_s \mathcal{F}_s$. For example, consider a grouping of two subgroups, $\mathcal{G} = \{s_1, s_2\}$, for $I = 10$ matched pairs. Suppose the first few pairs and the last pair make up the first subgroup and the rest are in the second subgroup, say $s_1 = \{1, 2, 3, 10\}$ and $s_2 = \{4, 5, 6, 7, 8, 9\}$. The set of potential responses, treatments, and covariates for the first group is then $\mathcal{F}_{s_1} = \{(r_{1s_1ij}^{(d_{1s_1ij})}, r_{0s_1ij}^{(d_{0s_1ij})}, d_{1s_1ij}, d_{0s_1ij}, \mathbf{X}_{s_1ij}, u_{s_1ij}) : s_1 = \{1, 2, 3, 10\}, i \in s_1, j = 1, 2\}$. The observed response, binary instrument, and exposure for a given unit in subset s_g is denoted as Z_{s_gij}, R_{s_gij}, and D_{s_gij} respectively.

2.2.2 Review: Matching, Instrumental Variables, and the CACE

Matching is a popular non-parametric technique in observational studies to balance the distribution of the observed covariates between treated and control units by grouping units based on the similarity of their covariates; see Stuart (2010), Chapters 3 and 8 of Rosenbaum (2010), and Rosenbaum (2020) for overviews of matching. Pair matching is a specific type of matching where each treated unit is only matched to one control unit. In the context of instrumental variables and pair matching, the instrument serves as the treatment/control variable and the matching algorithm creates I matched pairs where the two units in a matched pair are similar in their observed covariates x_{ij}, but one receives the instrument value $Z_{ij} = 1$ and the other receives the instrument value $Z_{ij} = 0$.

Instrumental variables (IV) is a popular approach to analyze causal effects when unmeasured confounding is present and is based on using a variable called an instrument (Angrist et al., 1996; Hernán and Robins, 2006; Baiocchi et al., 2014). The instrument must satisfy three core assumptions: (A1) the instrument is related to the exposure or treatment, or $\sum_{i=1}^{I} \sum_{j=1}^{2} (d_{1ij} - d_{0ij}) \neq 0$ (commonly referred to as instrument relevance); (A2) the instrument is not related to the outcome in any way except through the treatment, or $r_{0ij}^{(d)} = r_{1ij}^{(d)} \equiv r_{ij}^{(d)}$ for a fixed d (commonly referred to as the exclusion restriction); and (A3) the instrument is not related to any unmeasured confounders that affect the treatment and the outcome, or $P(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ within each pair i (commonly referred to as instrument ignorability or exchangeability). If these core assumptions are satisfied,

it is possible to obtain bounds on the average treatment effect (Balke and Pearl, 1997). To point identify a treatment effect, one needs to make additional assumptions. Here, we assume (A4) monotonicity where the potential treatment is a monotonic function of the instrument values, or $d_{0ij} \leq d_{1ij}$. Assumption (A4) can be interpreted in terms of four sub-populations: compliers, always-takers, never-takers, and defiers (Angrist et al., 1996). Compliers are units which their treatment values follow their instrument values, or $d_{0ij} = 0, d_{1ij} = 1$. Always-takers always take the treatment regardless of their instrument values, or $d_{0ij} = d_{1ij} = 1$. Never-takers never take the treatment regardless of their instrument values, or $d_{0ij} = d_{1ij} = 0$. Defiers act against their instrument values, or $d_{0ij} = 1, d_{1ij} = 0$. Assumption (A4) then states that no defiers exist.

Let N_{CO} be the total number of compliers in the population. Under the IV assumptions (A1)-(A4), the CACE, formally defined as

$$\lambda = \frac{\sum_{i=1}^{I}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(d_{1ij} = 1, d_{0ij} = 0)}{\sum_{i=1}^{I}\sum_{j=1}^{2} d_{1ij} - d_{0ij}} = \frac{1}{N_{\text{CO}}}\sum_{i=1}^{I}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(ij \text{ is a complier})$$

can be identified from data by taking the ratio of the estimated ITT effect over the estimated compliance rate. In the context of matching and instrumental variables, Baiocchi et al. (2010) and Kang et al. (2016a) proposed a test statistic to test the null $H_0 : \lambda = \lambda_0$ by using differences in the adjusted outcomes

$$T(\lambda_0) = \frac{2}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij}) \tag{2.1}$$

along with an estimator for the variance of $T(\lambda_0)$,

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\sum_{j=1}^{2}\left(Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij}) - T(\lambda_0)\right)^2 \tag{2.2}$$

Under the null, Baiocchi et al. (2010) and Kang et al. (2016a) showed that $\frac{T(\lambda_0)}{S(\lambda_0)}$ asymptotically follows a standard Normal distribution. For point estimation, the same set of authors proposed a Hodges-Lehmann type estimator (Hodges and Lehmann, 1963) which involves solving λ in the equation $T(\lambda)/S(\lambda) = 0$. For a $1 - \alpha$ % confidence interval, the equation $T(\lambda)/S(\lambda) \leq z_{1-\alpha/2}$ is solved for λ, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Normal distribution; see Kang et al. (2016a) and Kang et al. (2018) for details.

2.2.3 Heterogeneous Complier Average Causal Effect (H-CACE)

We formally define the target estimand of interest in the paper, the heterogeneous treatment effect among compliers, or H-CACE. Formally, the H-CACE is defined as the CACE for a subgroup of compliers with a specific value of covariates

$$\lambda(\mathbf{x}) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(d_{1ij}=1, d_{0ij}=0, \mathbf{X}_{ij}=\mathbf{x})}{\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{1ij}-d_{0ij})I(\mathbf{X}_{ij}=\mathbf{x})} \tag{2.3}$$

Because two units are assumed to have identical covariate values within each matched pair, $\lambda(\mathbf{x})$ can be rewritten as taking a subset of I matched pairs with identical covariates $\mathbf{x}$, say $s \subseteq \{1,\ldots,I\}$

$$\lambda_s = \frac{\sum_{i\in s}\sum_{j=1}^{2} r_{1sij}^{(d_{1sij})} - r_{0sij}^{(d_{0sij})}}{\sum_{i\in s}\sum_{j=1}^{2} d_{1sij} - d_{0sij}}$$

Since each H-CACE λ_s has the same form as the original CACE, we can apply the test statistic in Section 2.2.2. Formally, consider the subset-specific hypothesis $H_{0s}: \lambda_s = \lambda_0$ against $H_{1s}: \lambda_s \neq \lambda_0$. We can use the test statistic (2.1) with variance (2.2) among the pairs specific to subset s.

Also, under assumptions (A1)-(A4), for a mutually exclusive and exhaustive grouping $\mathcal{G} = \{s_1,\ldots,s_G\}$ of a set of pairs $s_g \subseteq \{1,\ldots,I\}$ with at least one complier within each subgroup s_g, the original CACE is equal to a weighted version of H-CACE:

$$\lambda = \sum_{g=1}^{G} w_{s_g}\lambda_{s_g}, \quad w_{s_g} = \frac{\sum_{i\in s_g}\sum_{j=1}^{2} d_{1sij} - d_{0sij}}{N_{\text{CO}}}.$$

An implication of this expression is that typical analysis of the CACE hides underlying effect heterogeneity. For example, suppose there are two subgroups defined by a binary covariate, say male or female, and consider two scenarios. In the first scenario, among compliers, 80% are male and 20% are female. Also, the H-CACE of male is 1.25 and the H-CACE of female is 0. In the second scenario, the male/female complier proportions remain the same, but the H-CACE of male is now 1.5 and the H-CACE of female is -1. In both scenarios, the CACE is 1. But, in the second scenario, females have a negative treatment effect. By only studying the CACE, as is typical in practice, variations in the treatment effects defined by H-CACEs would have been masked. The next section presents a way to unwrap the CACE and discover novel H-CACEs.

2.2.4 Discovering and Testing Novel H-CACE

A naive approach to finding and testing novel H-CACE would be to exhaustively test every H-CACE for every subset of matched pairs and gradually aggregate them based on their covariate similarities with appropriate statistical tests. However, this procedure will not only lead to false discoveries, but it will also be grossly underpowered.

Instead, based on the work by Hsu et al. (2015), we propose to use exploratory machine learning methods, such as CART, to discover and aggregate matched pairs into subgroups with similar treatment effects, formulating grouping $\mathcal{G}$. We will then use closed testing to test effect heterogeneity defined by these groups while strongly controlling the familywise error rate; see Algorithm 1 for details.

We explain in some detail the key steps in Algorithm 1. First, the specification of the null value λ_0 is for testing the sharp null of the form $H_0 : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$; this sharp null implies the "weak" or composite null $H_0 : \lambda = \lambda_0$ (Baiocchi et al., 2010). Setting $\lambda_0 = 0$ would test whether the H-CACE is zero or not and is the typical choice in most applications unless other null values are of scientific interest. Second, under the sharp null, the absolute value of the difference in adjusted outcomes between pairs, $\left|Y_i\right| = \left|(Z_{i1} - Z_{i2})(R_{i1} - \lambda_0 D_{i1} - (R_{i2} - \lambda_0 D_{i2}))\right|$, obscures the instrument assignment vector making $\left|Y_i\right|$ a function of $\mathcal{F}$ only, a fixed (and unknown) quantity. In contrast, Y_i is a function of both $\mathcal{F}$ and $\mathbf{Z}$. Consequently, conditional on $\mathcal{F}$, building a CART tree based on $|Y_i|$ as the response and $\mathbf{X}_i$ as the explanatory variables does not affect the distribution of $\mathbf{Z}$. The distribution of $\mathbf{Z}$ within each pair remains 1/2 as stated in assumption (A3) and is a key ingredient to achieve familywise error rate control for downstream inference; see our discussion on honest inference below.

Third, Algorithm 1 applies closed testing, a multiple inference procedure by Marcus et al. (1976), to test for multiple hypotheses about H-CACEs generated by CART's grouping $\mathcal{G} = \{s_1, \ldots, s_G\}$. Broadly speaking, closed testing will test sharp null hypotheses defined by every parent and child node of the estimated tree from CART and reject/accept these hypotheses while controlling for multiple testing issues; see Section 2.4.4 and Figure 2.5 for visualizations. A bit more formally, closed testing will test the global sharp null hypothesis $H_0 : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$ and subsequent subset-specific hypotheses $H_{0\mathcal{L}} : r_{1s_gij}^{(d_{1s_gij})} - r_{0s_gij}^{(d_{0s_gij})} = \lambda_0(d_{1s_gij} - d_{0s_gij})$ for all $g \in \mathcal{L}$, where $\mathcal{L}$ is a subset of the G groups formed by CART. We note that the difference between the global null and the subset-specific nulls is only in the pairs under consideration; all the nulls use the test statistics

Given : Observed outcome R, binary instrument Z, exposure D, covariates X, null value λ_0 for testing, and desired familywise error rate α

1 Pair match on observed covariates.

2 Calculate absolute value of pairwise differences for each matched pair

$$\left|Y_i\right| = \left|(Z_{i1} - Z_{i2})(R_{i1} - \lambda_0 D_{i1} - (R_{i2} - \lambda_0 D_{i2}))\right|$$

3 Construct mutually exclusive and exhaustive grouping using CART. Here, CART takes $|Y_i|$ as the outcome and $\mathbf{X}_i$ from each matched pair as the predictors. CART outputs a partition of covariates, which we use to define $\mathcal{G} = \{s_1, \ldots, s_G\}$ and consequently, H-CACEs.

4 Run closed testing (Marcus et al., 1976) to test statistical significance of H-CACEs for every subset $\mathcal{L} \subseteq \{1, \ldots, G\}$ of G groups where each subset defines the null hypothesis of the form $H_{0\mathcal{L}} : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$ for all $g \in \mathcal{L}$. Formally, run

for $\mathcal{L} \subseteq \{1, \ldots, G\}$ **do**
  if $H_{0\mathcal{L}}$ has not been accepted **then**
    Calculate $T_s(\lambda_0)$ and $S_s(\lambda_0)$ for $s = \bigcup_{g \in \mathcal{L}} s_g$
    if $\left|\frac{T_s(\lambda_0)}{S_s(\lambda_0)}\right| \leq z_{1-\alpha/2}$ **then**
      Accept the null hypothesis $H_{0\mathcal{K}} : \lambda_{\mathcal{K}} = \lambda_0$ for all $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, \mathcal{G}\}$
    end
    else
      Reject $H_{0\mathcal{L}}$
    end
  end
end

Output : Estimated and inferential quantities for H-CACEs (e.g. effect size, confidence interval, p-value) and novel H-CACEs from closed testing.

Algorithm 1: Proposed method to discover and test effect heterogeneity in IV with matching

introduced in Section 2.2.2. Also, the subset-specific hypotheses imply $H_{0\mathcal{L}} : \lambda_s = \lambda_0$ for $s = \bigcup_{g \in \mathcal{L}} s_g$. Closed testing would only reject the subset-specific hypotheses $H_{0\mathcal{L}}$ if all of the p-values from superset hypotheses $H_{0\mathcal{L}'}$, $\mathcal{L} \subseteq \mathcal{L}'$, are less than α.

As mentioned earlier, the key step of using $|Y_i|$ in CART allows for both discovery and downstream honest testing of H-CACEs via closed testing; again, honesty refers to control of the familywise error rate at level α when testing multiple hypotheses about H-CACEs that

were discovered by data. Because $|Y_i|$ is not a function of $\mathbf{Z}$, the original distribution of $\mathbf{Z}$ is preserved and we can use the standard randomization inference null distribution to honestly test each H-CACE discovered by CART. In fact, as noted in Hsu et al. (2015), this honesty property is preserved for any supervised machine learning algorithm that forms groups based on $\mathbf{X}$ and $|Y|$ as well as subsequent visual heuristics to check the algorithms' performance. Also, in recent work on estimating heterogeneous causal effects (Chernozhukov et al., 2018; Athey et al., 2019; Park and Kang, 2020), the notion of "honest" inference is often tied to sample splitting, where one subsample is used to discover different subgroups or to estimate nuisance parameters and the other subgroup is used to test the causal effect. Our approach does not have to use sample splitting to obtain honest inference and Proposition 2.1 shows this principle formally; Appendix A.2 shows this principle numerically.

Proposition 2.1 (Familywise Error Rate Control of Algorithm 1)**.** *Under the sharp null hypotheses $H_{0\mathcal{L}}$ in Algorithm 1, the conditional probability given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ that the algorithm makes at least one false rejection of the set of hypotheses is at most α.*

We now discuss some important limitations of Proposition 2.1 and the proposed algorithm. First, our algorithm's guarantee on controlling the familywise error rate is only for testing sharp nulls. As noted in Section 2, page 289 of Rosenbaum (2002a), testing for sharp nulls does not necessarily imply that the true data generating process always follow the sharp null and as such, the proposition makes no claims about how the true data generating process actually looks like. Having said that, the limitation of testing a sharp null versus a weak null has been discussed extensively; see Sections 3 and 4 of Rosenbaum (2002b), Ding (2017), Fogarty (2018), and Fogarty (2020). But, a recent work by Fogarty et al. (2021) has shown that testing the sharp null based on our test is an asymptotically valid test for the weak null; see Remark 1 of their Proposition 1. This suggests that the guarantees from Proposition 1 will likely hold even if we are testing weaker nulls with our algorithm. Second, a price we pay for using $|Y_i|$ to achieve honest inference is that we collapse the sign of the effect and therefore, CART treats subgroups with positive or negative effects equally. This is potentially problematic in settings where two different covariate values lead to identical effects (in magnitude), but different in signs; see Hsu et al. (2013) for additional discussions and Appendix A.4 for a numerical illustration. For our Medicaid example, if there is a partition of the covariates that leads to two identical H-CACEs in magnitude, but different in signs, our algorithm may not be able to detect the two subgroups. But, since using Medicaid is unlikely to be harmful, we don't believe this will be a significant concern

in our example, especially compared to the alternatives of not obtaining honest inference. Third, Proposition 2.1 does not describe the algorithm's statistical power to detect effect heterogeneity. The next section uses a simulation study to address power and other factors influencing discovery of H-CACEs.

2.3 Simulations

We conduct a simulation study to measure the performance of the proposed algorithm in two ways: (1) statistical power to test H-CACEs and (2) recovering effect modifiers. Throughout the simulation study, we vary the the compliance rate because prior works have shown that performance of IV methods depends heavily on the compliance rate, or more generally on the instrument's association to the treatment (i.e. instrument strength). In particular, problems can arise when the compliance rate is low; see Staiger and Stock (1997), Stock, Wright, and Yogo (2002), and references therein for more details.

Following Hsu et al. (2015), each simulation setting fixes the potential outcomes $r_{0ij}^{(d_{0ij})}$ and $r_{1ij}^{(d_{1ij})}$, potential treatments d_{0ij} and d_{1ij}, and covariates $\mathbf{X}_{ij}$ of each unit j within each of the $I = 2000$ pairs. There are six pre-instrument covariates, each generated from independent Bernoulli trials with 0.5 probability of success. At most two covariates, x_1 and x_2, modify the treatment effect. That is, H-CACEs defined by $\lambda(x_1, \ldots, x_6)$ in equation (2.3) depend on at most two covariates, x_1 and x_2. Also, because both x_1 and x_2 are binary, there are at most four different H-CACEs defined by different combinations of binary variables λ_{00}, λ_{01}, λ_{10}, and λ_{11}; for notational simplicity, we use $\lambda_{x_1x_2}$ to represent equation (2.3). Similar to the design of the OHIE, the data is generated under the assumption of one-sided compliance. This means that for every unit, the potential treatment having not received the instrument is 0, $d_{0ij} = 0$. The potential treatment having received the instrument, d_{1ij}, is then a Bernoulli trial with success rate π; π is also the compliance rate. In Appendix A.3, we consider the setting in which the compliance rate may depend on x_1 and x_2, say via $\pi_{x_1x_2}$. Finally, the potential outcomes having not received the instrument $r_{0ij}^{(d_{0ij})}$ are from a standard normal distribution $r_{0ij}^{(d_{0ij})} \sim N(0, 1)$, and the potential outcomes having received the instrument $r_{1ij}^{(d_{1ij})}$ are a function of the H-CACE $r_{1ij}^{(d_{1ij})} = r_{0ij}^{(d_{0ij})} + d_{1ij}\lambda_{x_1x_2}$. Once all the potential treatment and outcomes are generated, the observed treatment and outcome are determined based on the value of the instrument and SUTVA. Finally, the regression tree in Algorithm 1 is estimated in R using the package *rpart*, version 4.1-15 (Therneau,

Atkinson, and Ripley, 2015). Unless specified otherwise, we use a complexity parameter of 0.005 (half of the default setting) and use defaults for the rest of *rpart*'s parameters. Our proposed method is referred to as "H-CACE" in the results below.

For comparison, we also apply a recent method by Bargagli-Stoffi et al. (2019) to discover and test H-CACEs. Briefly, their method, which we refer to as "BCF-IV" in the results below, utilizes modern tree-based methods (Athey and Imbens, 2015; Hahn et al., 2020) to estimate heterogeneous intent-to-treat (ITT) effects and suggests different sub-populations of interest; we remark that unlike our proposal, their method does not use matching and uses the original, untransformed R_{ij} inside the tree fitting step. Then, for each sub-population, the method estimates and tests its H-CACE using the two-stage least square estimator. We use the *bcf_iv* function available on the authors' Github repository and use the default parameters of *rpart* and *bcf* (Hahn et al., 2020).

2.3.1 Statistical Power

To measure a method's statistical power when the subgroup-specific null hypotheses aren't specified a priori, we divide the number of false null hypotheses rejected by the total number of false null hypotheses suggested by the method. We refer to this rate as the true discovery rate; note that if the number of false nulls being suggested is fixed, the true discovery rate is one minus the proportion of false null hypotheses retained.

We compute the true discovery rate at varying levels of instrument strength and four heterogeneous treatment settings: (a) No Heterogeneity, (b) Slight Heterogeneity, (c) Strong Heterogeneity, and (d) Complex Heterogeneity. In setting (a), there are no effect modifiers resulting in one subgroup with equal treatment effects, $\lambda_{00} = \lambda_{01} = \lambda_{10} = \lambda_{11} = 0.5$. In setting (b), there is one effect modifier x_1 resulting in two subgroups with similar but different treatment effects, $\lambda_{00} = \lambda_{01} = 0.7$ and $\lambda_{10} = \lambda_{11} = 0.3$. In setting (c), there is one effect modifier x_1 resulting in two subgroups with dissimilar treatment effects, $\lambda_{00} = \lambda_{01} = 0.9$ and $\lambda_{10} = \lambda_{11} = 0.1$. And, in setting (d), there are two effect modifiers x_1 and x_2 resulting in three subgroups, one with a strong effect, two with no effects, and the last group with the average effect, $\lambda_{00} = 1.5$, $\lambda_{01} = \lambda_{10} = 0$ and $\lambda_{11} = 0.5$. In all four settings, the overall complier average causal effect is $\lambda = 0.5$.

We repeat the simulation 1000 times for each treatment heterogeneity and instrument strength combination. We remark that the null hypothesis is that of no treatment effect (i.e. $\lambda_0 = 0$) and only the hypotheses consisting of pairs with $\lambda_{x_1 x_2} = 0$ are true null hypotheses.

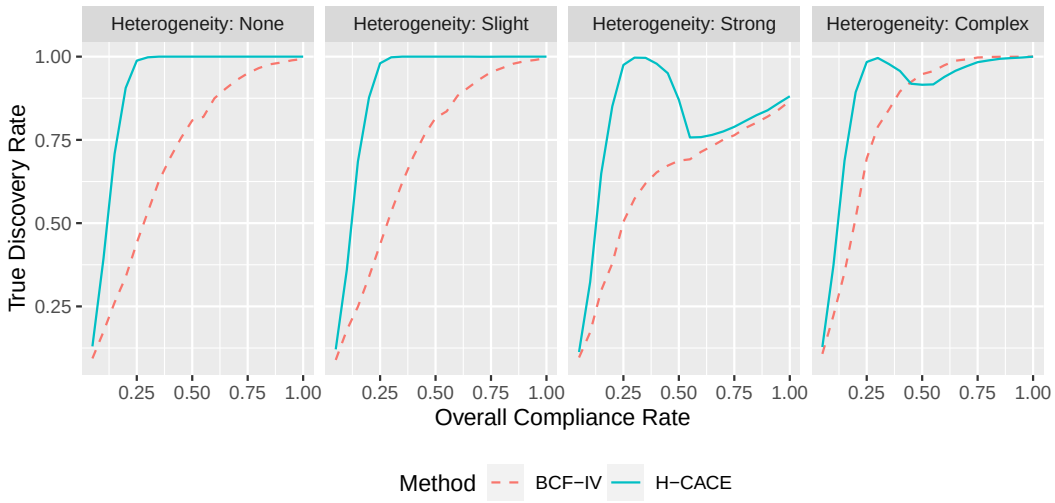


Figure 2.1: True discovery rate as a function of the compliance rate and heterogeneity settings. The dashed and solid lines denote the BCF-IV procedure and our proposed algorithm, respectively.

Figure 2.1 shows the true discovery rate under four treatment heterogeneity settings. We see that as the compliance rate (i.e. instrument strength) increases, the true discovery rate of our method grows across all settings. In particular, our approach has the best power in the region where the compliance rate is low, roughly under 40%. Even when the compliance rate is high, we see that BCF-IV generally has lower power than our method across different heterogeneity settings, especially in the No Heterogeneity, Slight Heterogeneity, and Strong Heterogeneity settings. In the complex heterogeneity setting, we see the true discovery rate is rather similar between the two methods. This is because this setting has the largest discrepancies in H-CACEs between subgroups, and thus, it is easy for CART to correctly split on the covariates x_1 and x_2. Further, with the large magnitudes of H-CACEs in this setting, the null hypothesis tests are more easily rejected in favor of the alternative.

We also take a moment to explain a counter-intuitive dip in our method's true discovery rate under the strong and complex heterogeneity settings in Figure 2.1. Briefly, this drop in the true discovery rate is due to the formation of leaves with smaller treatment effects. As the compliance rate becomes large, these small effects begin to be get suggested by CART. But, the power to reject the null in favor of these small effects are small and the overall true discovery rate dips briefly. However, as the compliance rate reaches one, we see the

true discovery rate of our method begin to climb again; Appendix A.5 contains additional details surrounding this phenomena.

Another interpretation of this dip reflects a limitation of the true discovery rate as a metric for statistical power in certain settings of multiple testing with hypotheses adaptively generated by tree-based algorithms. Specifically, because the true discovery rate only considers hypotheses generated by the tree, if a tree were to not split (or rarely split) and the overall effect is strong, there would only be one hypothesis in the denominator of the true discovery rate and the lone hypothesis would likely be rejected, leading to a true discovery rate of one. The Strong and Complex Heterogeneity settings under low compliance rates, especially before the dip, is a reflection of this phenomena where our tree fails to split, resulting in one single hypothesis that is eventually rejected; see Appendix A.5 for additional discussions. Nevertheless, as Figure 2.1 shows, our method consistently shows a higher true discovery rate compared to BCF-IV across many settings and our method is more likely to discover true non-zero effects than BCF-IV.

2.3.2 False Positive Rate and F-Score

We also assess our algorithm's ability to predict effect modifiers from $\mathbf{X}_{ij}$. Specifically, we say that a method predicts a variable to be an effect modifier when the tree splits on the variable and rejects one of the hypotheses of the split's children. In contrast, if either (a) the tree splits on a variable, but none of the hypotheses defined by the split is rejected or (b) the tree does not split on the variable, the variable is not predicted to be an effect modifier. For example, for a given tree that splits only on covariate x_1, if at least one of the subgroup-specific null hypotheses is rejected, x_1 is predicted to be an effect modifier. Instead, if none of the subgroup-specific null hypotheses are rejected then x_1, as well as other variables not selected by the tree, are not predicted to be effect modifiers. We then use the F-score and the false positive rate (FPR) common in the classification literature to measure a method's ability to correctly predict effect modifiers. The F-score is the harmonic mean of recall and precision, or alternatively,

$$F = \frac{TP}{TP + 0.5(FP + FN)}$$

where TP stands for true positives, FP stands for false positives, and FN stands for false negatives; see Table 2.1 for details. The F-score ranges from zero to one with a value closer

to one implying greater accuracy. The FPR is defined as $FPR = FP/(FP + TN)$ and ranges from zero to one, with a value close to zero being preferred.

	True Condition	
Method's Prediction	Variable is an effect modifier	Variable is not an effect modifier
Predicted as effect modifier	True Positive (TP)	False Positive (FP)
Not predicted as effect modifier	False Negative (FN)	True Negative (TN)

Table 2.1: Binary classification table for effect modifiers.

We use the same four heterogeneity settings of (a) No Heterogeneity, (b) Slight Heterogeneity, (c) Strong Heterogeneity, and (d) Complex Heterogeneity. Figure 2.2 shows the results of the F-score and FPR from our proposed algorithm and BCF-IV. Across four settings, our proposal has a false positive rate of nearly zero, never falsely declaring a variable to be an effect modifier. In contrast, BCF-IV has a larger false positive rate, declaring variables to be effect modifiers when they do not actually modify the compliers' effect. For example, in setting (a) without any effect modifiers, BCF-IV has a false positive rate hovering above 50% whereas our method has a false positive rate of 0%. In other words, BCF-IV falsely declared at least one of the six covariates as an effect modifier roughly 50% of the time whereas our method never declared any of the six covariates as effect modifiers.

However, our algorithm's F-score is generally smaller than that from BCF-IV unless the compliance rate is high and the effect heterogeneity is strong. In particular, when the compliance rate is roughly under 50% or if two subgroups have similar effect sizes, our method cannot predict the effect modifiers as well as BCF-IV. But, when the compliance rate is above 50% and the effect heterogeneity is strong, our algorithm has a similar F-score as BCF-IV. Overall, the low F-score is a price that our algorithm pays for making sure that the FPR is small. In contrast, BCF-IV has a higher F-score, but pays a price with a high FPR.

In the supplementary materials Appendices A.3, A.4, A.6, we conduct additional simulation studies where we (i) vary the compliance rate by covariates, (ii) allow H-CACEs to be equal in magnitude, but opposite in direction to measure the effect of using $|Y_i|$ in our algorithm, and (iii) demonstrate the two methods in a simulation that closely resembles the data from the OHIE, where there are more than two effect modifiers. To summarize the results, for (i) and (iii), the story is very similar to what's presented here, where our method has high true discovery rate, low FPR and F-score compared to those from BCF-IV. For

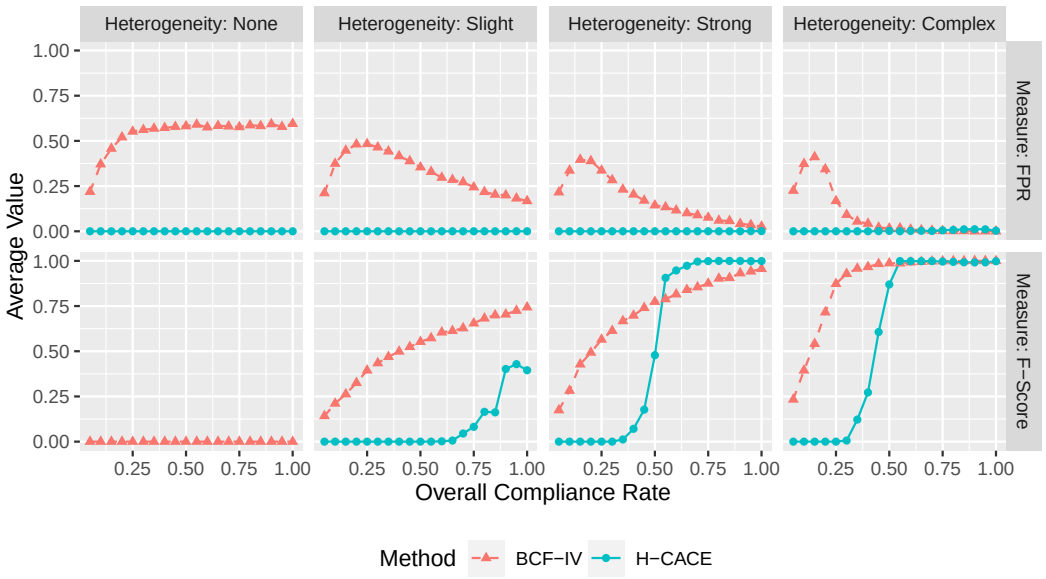


Figure 2.2: F-score and false positive rate as a function of compliance rate and heterogeneity settings. The solid lines with circles denote our proposed algorithm and the dashed lines with triangles denote BCF-IV.

(ii), as expected, we find that our method has a low true discovery rate, FPR, and F-score. But, as soon as the magnitudes of the H-CACEs are dissimilar, our method returns to the case presented here.

2.3.3 Takeaways from the Simulation Study

Overall, the simulation study shows that our algorithm has large statistical power and low false positive rates across all settings. In contrast, the BCF-IV algorithm has low power and produces large FPRs, especially when no effect heterogeneity exists in the data; in other words, BCF-IV often falsely declares a variable to be an effect modifier. But, our algorithm generally has a low F-score compared to that from BCF-IV except in regimes where the effect heterogeneity is strong and the compliance rate is high.

We remark that the simulations studies do not encapsulate every type of effect heterogeneity and it is possible that our method may suffer in certain settings. In particular, as discussed above, because our method tends to be conservative in predicting effect modifiers in order to guard against discovering spurious heterogeneity, we suspect that if there

are many effect modifiers compared to spurious effect modifiers, our method may not be able to detect all of the effect modifiers. This suspected degradation in performance was not observed when we had 5 effect modifiers among 15 potential effect modifiers in our simulation study that mimicked the data from the OHIE. But, further investigation is warranted, especially if the number of effect modifiers and/or the number of covariates is high dimensional.

We also remark that the simulation results in Sections 2.3.1 and 2.3.2 do not necessarily contradict each other. Roughly speaking, the result in Section 2.3.1 concerns the ability for algorithms to have high *statistical power* whereas the result in Section 2.3.2 concerns the ability for algorithms to *predict* variables. An algorithm like BCF-IV could liberally predict many effect modifiers, generally leading to a high F-score, but a high FPR. Also, the power to test the nulls suggested by the predicted effect modifiers could be low since the selected variables will define many (likely small) subgroups. In contrast, an algorithm like ours could conservatively predict effect modifiers, leading to a small F-score, but a low FPR. Also, the power to test the nulls suggested by the predicted variables could be high since most of the selected variables will be effect modifiers. In short, our method is somewhat cautious, but certain whereas BCF-IV is optimistic, but somewhat error-prone.

2.4 Analysis of the Oregon Health Insurance Experiment

2.4.1 Data Description

We use our method to analyze the heterogeneous effects of Medicaid on the number of days an individual's physical or mental health prevented their usual activities in the past month. In brief, the OHIE collected administrative data on hospital discharges, credit reports, and mortality, survey data on health care utilization, financial strain, and overall health, and pre-randomization demographic data. There were 11,808 lottery winners and 11,933 lottery losers in the publicly available survey data for a total sample size of 23,741 individuals; see Finkelstein et al. (2012) for details.

We matched on the following demographic, pre-randomization variables recorded by Finkelstein et al. (2012): sex, age, whether they preferred English materials when signing up for the lottery, whether they lived in a metropolitan statistical area (MSA), their education level (less than high school, high school diploma or General Educational Development (GED), vocational or 2-year degree, 4-year college degree or more), and self-identified race

(as the individual reported in the survey). Since some of the covariates had missing data, namely self-identifying as Hispanic or Black and their level of education, we also matched on indicators of their missingness; see Section 9.4 of Rosenbaum (2010) for details. We used the R package *bigmatch*, version 0.6.1, (Yu, 2019) with an optimal caliper and a robust rank-based Mahalanobis distance to generate our optimal pair match. Figure 2.3 shows covariate balance before and after matching.

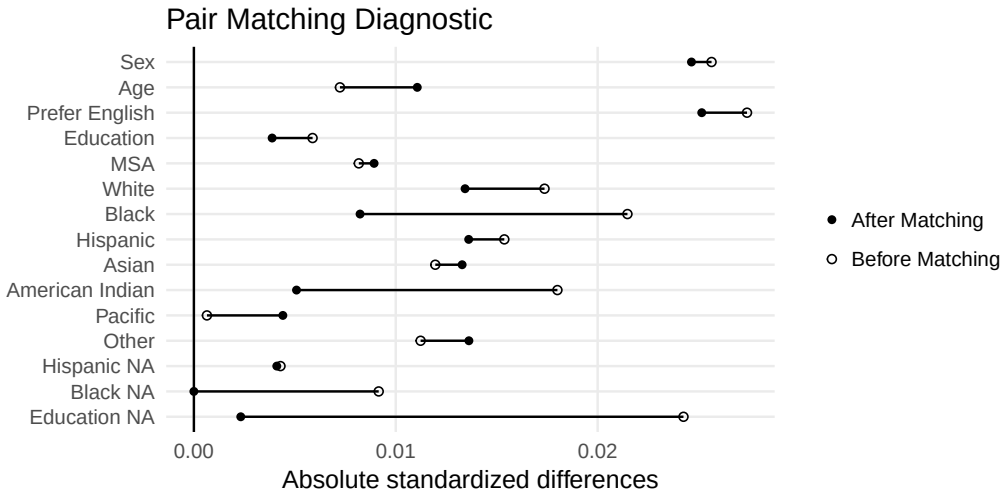


Figure 2.3: Covariate balance as measured by difference in means of the covariates between the treated and control groups, before and after matching.

For the majority of covariates, the matching algorithm did little to change the absolute standard differences between lottery winners and losers. This is not surprising given that the lottery was randomized. However, the indicator for missingness in education, self-identified American Indian, and Black were made to be more similar after matching. An absolute standardized difference of 0.25 is deemed acceptable (Rubin, 2001; Stuart, 2010), which our covariates satisfied after matching.

2.4.2 Instrument Validity

Before we present the results of our analysis using the proposed method, we discuss the plausibility of the lottery as an instrument. The lottery is randomized which ensures that the instrument is unrelated to unmeasured confounders and satisfying (A3). Winning the lottery, on average, increased enrollment of Medicaid by 30% (Finkelstein et al., 2012),

satisfying (A1). Assumption (A4) in the context of the OHIE states that there are no individuals who defy the lottery assignment to take (or not take) Medicaid if they lost (or won) the lottery. This is guaranteed by the design of the lottery, since an individual who lost the lottery cannot have access to Medicaid. However, we remark that Finkelstein et al. (2012) measured the treatment as whether or not an individual has ever had Medicaid during the study and a few individuals were already enrolled in Medicaid before the lottery winners were announced. Finally, assumption (A2) is the only assumption that could potentially be violated since individuals were not blind to their lottery results. This theoretically allowed lottery losers to seek other health insurance or lottery winners to make less healthy decisions since they're now able to be insured. These changes in an individual's behavior could affect his/her outcome regardless of his/her treatment and thus, may violate (A2).

2.4.3 Analysis and Results

We run Algorithm 1 and present the results in Figure 2.4. We remark that we used *rpart* in R with a complexity parameter of 0 and maximum depth of 4. The depth of the tree was chosen by forming trees of larger depth and then pruning back until a more interpretable tree was obtained. For each node of the CART, we tested whether or not there is an effect of enrolling in Medicaid $H_{0s} : \lambda_s = 0$. In Figure 2.4, a solid lined box denotes a null hypothesis that was rejected and a dashed lined box denotes a null hypothesis that was retained, both by the closed testing procedure. Each node contains its estimated H-CACE $\hat{\lambda}_s$, 95% confidence interval, the number of pairs I_s, and the estimated compliance rate $\hat{\pi}_s$. Here, a positive H-CACE implies a decrease in the number of days where the individual's physical and mental health prevented them from their usual activities, and a negative value implies an increase; in short, positive effects are beneficial to individuals. Also, some nodes imply a significant effect of Medicaid at level 0.05, but are enclosed in a dashed lined box. This is due to the closed testing procedure; an intersection of hypotheses containing the node in question was not rejected, and so any hypotheses in this intersection could not be rejected.

From Figure 2.4, we can see evidence of heterogeneous treatment effects among the complier population. Specifically, Medicaid had a strong effect (1) among complying non-Asian men over the age of 36 and who prefer English, as well as (2) complying individuals younger than 36, who prefer English, and do not have more than a high school diploma or GED. Interestingly, among non-Asians over the age of 36 and who prefer English, females

did not benefit from Medicaid as much as males even though the female subgroup was larger than the male subgroup and the compliance rates between the two subgroups were similar.

More generally, while there is some variation in the compliance rates between groups, most of them are minor and hover between 25% to 30%. The minor variation suggests that while some subgroups are more likely to be compliers than others, most of the effect heterogeneity is likely driven by the variation in how the treatment differentially changes the response across subgroups; a bit more formally, most of the effect heterogeneity is likely arising from the numerator of the H-CACE rather than the denominator of the H-CACE.

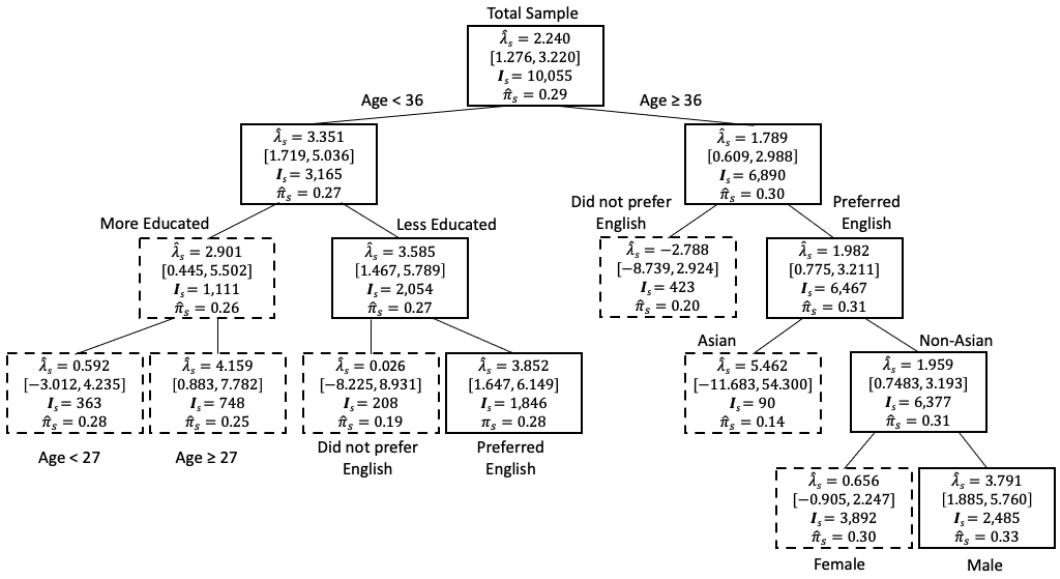
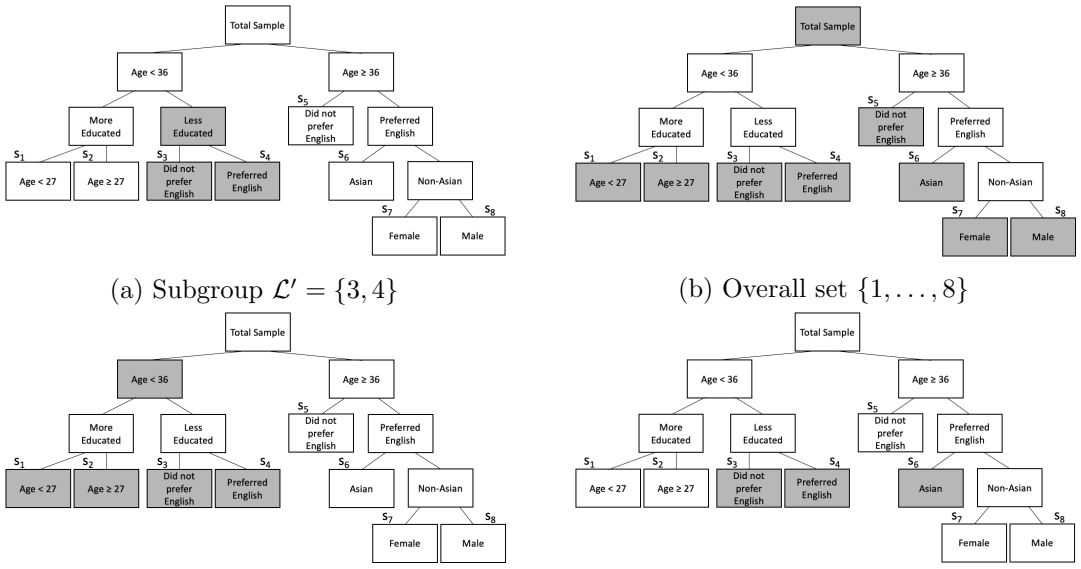


Figure 2.4: Results of our proposed method on the effect of enrolling in Medicaid on the number of days physical or mental health did not prevent usual activities. Here, less educated refers to pairs with at most a high school diploma or GED and more educated refers to pairs with a higher education. Also, positive effects are beneficial to individuals. Solid lined boxes denote hypothesis tests that were rejected and dashed lined boxes denote hypotheses that were retained by closed testing. Within each box, the subgroup-specific estimated H-CACE $\hat{\lambda}_s$, its 95% confidence interval, sample size of pairs I_s, and the estimated compliance rate $\hat{\pi}_s$ are provided.



(a) Subgroup $\mathcal{L}' = \{3, 4\}$

(b) Overall set $\{1, \ldots, 8\}$

(c) Set $\{1, \ldots, 4\}$

(d) Set $\{3, 4, 6\}$

Figure 2.5: Illustration of closed testing to test the null hypothesis H_{0s_4} for all $j = 1, 2$ and $i \in s_4$. Each subplot highlights subsets required to be tested and rejected as part of closed testing.

2.4.4 An Example of Closed Testing

To better illustrate the closed testing portion of Algorithm 1, we walk through an example of the testing procedure based on the OHIE. As seen in Figure 2.4, CART produced a tree with $G = 8$ leaves. Now, consider testing whether there is evidence of a heterogeneous effect of Medicaid for young individuals who prefer English and have at most a high school diploma or GED, i.e. node s_4 in Figure 2.5 and $\mathcal{L} = \{4\}$ using Algorithm 1's notation. The null hypothesis of interest would be H_{0s_4}, for all $j = 1, 2$ and $i \in s_4$. We then test and reject all of the hypothesis tests containing group s_4. For example, we need to test the null hypothesis concerning the ancestor of s_4, say the subgroup of individuals who are younger than 36 and have at most a high school diploma or GED denoted as $\mathcal{L}' = \{3, 4\}$; see part (a) of Figure 2.5. Additionally, we need to test and reject all of the supersets containing $\mathcal{L}'$, which include but are not limited to the overall set $\{1, \ldots, 8\}$, $\{1, \ldots, 4\}$, and $\{3, 4, 6\}$. If every superset hypothesis and H_{0s_4} are rejected at level α, we can declare the effect in node s_4 to be significant and, by Proposition 2.1, the familywise error rate is controlled at α. Repeating this process for every node in the tree will give the results in Figure 2.4.

2.5 Discussion

In this paper, we propose a method based on matching to detect effect heterogeneity using an instrument. Under the usual IV assumptions, our method discovers and tests heterogeneity in the complier average treatment effect by combining matching, CART, and closed testing, all without the need to do sample splitting. The latter is achieved by taking the absolute value of the adjusted pairwise differences to conceal the instrument assignment and this allows our proposed method to control the familywise error rate. We also conducted a simulation study to examine the performance of our method and compared it to a recent method referred to as BCF-IV. Our method was then used to study the effect of Medicaid on the number of days an individual's physical or mental health did not prevent their usual activities where we used the lottery selection as an instrument. We found that Medicaid benefited complying, older, non-Asian men who selected English materials at lottery sign-up and for complying, younger, less educated individuals who selected English materials at lottery sign-up.

We conclude by making some recommendations about how to properly use our algorithm in practice, especially in light of existing approaches. First, as explained in the introduction, when there is noncompliance, exploring heterogeneity in the ITT alone with existing methods may provide an incomplete picture of the nature of the treatment effect. Relatedly, in settings where unmeasured confounding is unavoidable, our method based on an instrument is a promising way to discover and test effect heterogeneity.

Second, as alluded to in Section 2.3.3, the simulation results suggest that our algorithm tends to be conservative in discovering novel effect modifiers, reporting effect modifiers only if there is strong evidence for heterogeneity and minimizing prediction of spurious effect modifiers. In other words, investigators can be reasonably confident that effect heterogeneity exists among the variables declared by our algorithm as "real" effect modifiers. But, those variables that are not predicted by our algorithm may also be true effect modifiers, and in such cases, investigators may need additional samples to detect them using our method. In contrast, BCF-IV tends to be anti-conservative, reporting more effect modifiers, some of which may be spurious effect modifiers. While this may be advantageous in situations where there is slight effect heterogeneity or where exploration for effect heterogeneity is encouraged, investigators may not feel as confident about whether the detected effect heterogeneity truly exists.

Third, how our method performs in settings with potentially high dimensional effect

modifiers is not fully understood. In particular, while our method performed well when the structure of effect heterogeneity grew more complex, or when the number of effect modifiers were 5 out of 15 potential effect modifiers, the simulations did not consider the setting of moderate to high dimensional effect modifiers, and future research is warranted.

Fourth, most recent approaches on effect heterogeneity, notably Chernozhukov et al. (2018), utilize sample splitting to achieve honest inference (i.e. type I error rate control) whereas our method uses absolute value of matched pairs to achieve it; note that both methods theoretically allow for a large class of machine learning methods to detect heterogeneous treatment effects, even though ours focused on CART for its simplicity and interpretability. While our method uses the full sample for both discovery and honest testing compared to those based on sample splitting, one of the caveats of our method is that our method may not be able to detect subgroups with identical effect sizes, but in opposite signs. Overall, every algorithm for effect heterogeneity carries some trade-offs and we urge investigators to understand their strengths and limitations to solidify and strengthen causal conclusions about effect heterogeneity in IV studies.

3 INDIVIDUALIZED TREATMENT RULES WITH MULTILEVEL INSTRUMENTAL VARIABLES

3.1 Introduction

Choosing between competing treatments for different patients can be a challenging endeavor, particularly when the different patients exhibit different demographics, health statuses, and medical histories. As investigators and practitioners work to decide which treatment would most benefit a patient, it may be advantageous to estimate treatment effects specific to subgroups. The subgroup-specific treatment effects can then be used to develop individualized treatment strategies to provide the treatment most appropriate for the patient. Such treatment strategies are referred to as individualized treatment rules (ITR), or optimal decision rules, and are typically used to optimize the value function, or the expected mean outcome over a target population.

Estimation of optimal ITR has been extensively studied, where methods generally fall within two approaches, indirect and direct. Indirect methods implement a two-step process using a regression model. First, the outcome is regressed on the treatment, covariates, and treatment-covariates interactions. Second, the treatment that optimizes the predicted mean outcome for given covariate values is chosen (e.g. Zhao et al. (2009); Qian and Murphy (2011); Moodie et al. (2014); Sutton and Barto (2018)). Direct methods reformulate the problem as a weighted misclassification problem, so as to avoid the need of modeling the outcome. The direct method introduced by Zhao et al. (2012), referred to as Outcome Weighted Learning (OWL), has been inspirational for further weighting techniques estimating optimal ITR to better accommodate different data settings (e.g. Zhang et al. (2012); Xu et al. (2015); Chen et al. (2016), Chen et al. (2017); Lou et al. (2018); Chen et al. (2018)). However, both indirect and direct methods require the assumption of ignorability, that there is no unmeasured confounding on the effect of the treatment on the outcome. Unfortunately, this condition may not hold in observational studies or randomized trials with noncompliance.

Valid instrumental variables (IV) are variables related to the outcome only through the treatment and can be used to infer treatment effects in the presence of unmeasured confounding. Under some core IV assumptions, bounds on the treatment effect can be obtained [Balke and Pearl (1997)]. To point identify a treatment effect, an additional point identification assumption is used, namely monotonicity [Angrist et al. (1996)], or there is

no interaction between the instrument and unmeasured confounder on the additive effects of the instrument on the treatment [Wang and Tchetgen Tchetgen (2018)]. Some IVs commonly used in health studies, many of which are not naturally binary, are assignment to a treatment in randomized trials with noncompliance, encouragement to take a treatment, distance to specialty care providers, physician or hospital preference, multiple genetic variants, timing of admission, and insurance plans; for additional discussions and review, see Hernán and Robins (2006) and Baiocchi et al. (2014). There has been much progress in estimation and inference of treatment effects using IV, mostly using likelihood, series, sieve, minimum distance, matching, and/or moment-based methods; see Abadie (2003); Blundell and Powell (2003); Newey and Powell (2003); Ai and Chen (2003); Hall and Horowitz (2005); Tan (2006); Blundell et al. (2007); Baiocchi et al. (2010); Kang et al. (2016a); Darolles et al. (2011); Chen and Pouzo (2012); Okui et al. (2012); Su et al. (2013); Athey et al. (2019) and references therein. However, work incorporating IV estimation to derive optimal ITR is still understudied.

Only recently have IV techniques been used to identify optimal ITR. An extension of OWL to utilize a valid instrument was proposed by Cui and Tchetgen Tchetgen (2021b), who also provide multiply robust estimators for ITR identification. Qiu et al. (2021) exploit an IV to estimate optimal ITR when the treatment is a limited resource. Pu and Zhang (2020) estimate optimal ITR with an IV using only bounds on the treatment effect, when point identification assumptions may not necessarily hold. A necessary and sufficient condition for identifying optimal ITR using an IV was first proposed in Cui and Tchetgen Tchetgen (2020) and later discussed in Cui and Tchetgen Tchetgen (2021a). Estimating dynamic treatment rules, individualized treatment rules that may change over time, using an IV are presented by Chen and Zhang (2021). For further discussions, see Cui and Tchetgen Tchetgen (2021a). However, all of the previous work consider the use of a binary instrument to identify an optimal ITR. As many of the IVs commonly used are not binary, treating them as such is a limitation in practice.

Here, we introduce methods to identify optimal ITR in the presence of unmeasured confounding using multiple, discrete valued instruments, or equivalently, multilevel instruments. Specifically, we propose both indirect and direct methods to identify optimal ITR using a multilevel IV for either the overall population or the subpopulation of compliers, who are patients who comply with their assigned treatment. The methods proposed here generalize techniques identifying optimal ITR using a binary IV to multilevel IV through the use of contrasts, or linear combinations of variables with coefficients that sum to zero.

When the instrument is binary, our proposed methods reduce to existing methods using instruments to identify optimal ITR. As part of the extension of techniques using binary IV, we propose methods using both observed covariates and latent effect modifiers to identify optimal ITR, and methods individualizing the estimation of ITR. Additionally, we extend the necessary and sufficient condition for identifying optimal ITR using a binary IV [Cui and Tchetgen Tchetgen (2020)] to multilevel instruments and use it to bridge the setting of identifying optimal ITR for the overall population with the setting of identifying optimal ITR for the complier subpopulation. Extensive simulations are conducted to evaluate the performance of our proposed methods for both the overall population and complier subpopulation under varying strengths of a multilevel instrument. We then apply our methods to an observational study comparing the effects of two competing treatments for carotid-artery disease, carotid endarterectomy versus carotid artery stenting, on preventing stroke or death within 30 days of their index procedure.

3.2 Methodology

3.2.1 Notation

Let Y denote the outcome of interest. Without loss of generality, we assume that a larger outcome is preferred. The observed binary treatment $A \in \{-1, 1\}$ is allowed to be confounded in its effect on Y by an unmeasured variable U, so that a naive estimate of the treatment effect of A on Y would be biased. The discrete instrument is denoted as Z and has support $\{1, \ldots, k\}$, where k is the number of levels of the IV. We let $X \in \mathcal{X}$ denote the observed pre-instrument covariates, where $\mathcal{X}$ is a p-dimensional vector. The full data $(Z_i, X_i, A_i, U_i, Y_i)$, for $i = 1, \ldots, n$ are assumed to be identically and independently distributed across the n observations. As U is unobserved, the observed data is (Z_i, X_i, A_i, Y_i).

To discuss the necessary assumptions and estimate the causal effects used to identify optimal ITR, we adopt the potential outcomes framework in Neyman (1923) and Rubin (1974). The potential outcomes for a given instrument value $Z_i = z$ and treatment $A_i = a$ are denoted $Y_i(z, a)$, and the potential treatments for a given instrument value $Z_i = z$ are denoted $A_i(z)$. We assume positivity for the instrument, that each individual can receive any value of the discrete instrument ($0 < P(Z = z|X) < 1$ almost surely) and Stable Unit Treatment Value Assumption (SUTVA) on both the outcome and treatment for the

instrument. SUTVA here states that there is no interference among individuals, there is only one version of the instrument levels, and we observe the potential treatment $A_i(z)$ and potential outcome $Y_i(z, A_i(z))$ if individual i receives instrument value $Z_i = z$ [Rubin (1980)]. SUTVA can be written as $A_i = \sum_{z=1}^{k} 1\{Z_i = z\}A_i(z)$ and $Y_i = \sum_{z=1}^{k} 1\{Z_i = z\}Y_i\big(z, A_i(z)\big)$. We assume SUTVA and positivity on the instrument throughout this work, and therefore do not explicitly refer to them when stating our results.

3.2.2 Review: Iindividualized Treatment Rules and Instrumental Variables

Under the potential outcome framework, the optimal decision rule for individual i, denoted $d_i \in \{-1, 1\}$, can be clearly defined as $d_i = \text{sign}\{Y_i(1) - Y_i(-1)\}$, where $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = -1$ if $x < 0$, since we assume larger outcomes are preferred. As only one of the potential outcomes, $Y_i(1)$ or $Y_i(-1)$, can be observed [Holland (1986)], we must use causal identification assumptions, SUTVA, positivity, and ignorability on the treatment, so that there is no unmeasured variable U confounding the effect of the treatment, to identify the average potential outcome $E[Y(a)] = E[Y|A = a]$. Under the assumptions SUTVA, positivity, and ignorability on the treatment, we can further use the conditional average treatment effects to define and identify the optimal ITR,

$$d^*(X_i) = \text{sign}\{E[Y_i(1) - Y_i(-1)|X_i]\} = \text{sign}\{E[Y_i|A_i = 1, X_i] - E[Y_i|A_i = -1, X_i]\}.$$

This motivates indirect methods, as the sign of the conditional treatment effects is the treatment optimizing the predicted mean outcome when larger outcomes are preferred. Assuming SUTVA on the treatment, $Y_i = \sum_a 1\{A_i = a\}Y_i(a)$, the optimal treatment rule can be rewritten as $d^*(X) = \arg\max_{d \in \mathcal{D}} E[Y(d(X))]$, where $\mathcal{D}$ is the set of allowable treatments and $Y(d(X))$ denotes the potential outcome of being assigned decision $d(X)$. The optimal decision rule $d^*(X)$ then satisfies $E[Y(d(X))] \leq E[Y(d^*(X))]$ for all $d \in \mathcal{D}$ and maximizes the expected population outcome, or value function, $V(d(X)) = E[Y(d(X))]$.

Alternatively, the identification of an ITR can be converted to a weighted classification problem and solved directly by OWL methods [Qian and Murphy (2011); Zhang et al. (2012); Zhao et al. (2012); Xu et al. (2015); Chen et al. (2016), Chen et al. (2017); Zhou and Kosorok (2017); Lou et al. (2018)]. Leveraging inverse probability weighting, the optimal

individualized treatment rule can be identified by directly maximizing the value function

$$d^*(X) = \arg\max_{d\in\mathcal{D}} E[Y(d(X))] = \arg\max_{d\in\mathcal{D}} E\left[\frac{\mathbb{1}\{A = \mathcal{D}(X)\}Y}{P(A=1|X)}\right].$$

In practice, the optimal decision rule is estimated by minimizing the empirical value of the weighted misclassification error $E\left[\frac{Y}{P(A=1|X)}\mathbb{1}\{A \neq \text{sign}(f(X_i))\}\right]$ over a prespecified set of functions, since we can represent the set of decisions as $\mathcal{D} = \text{sign}\{f(x)\}$. To overcome the nonconvexity of the 0-1 loss, a surrogate loss function, such as the hinge loss, is used in its place.

In observational studies or in randomized trials with noncompliance, due to unmeasured confounding, the expectation of the outcome conditioned on the treatment and observed covariates no longer identifies the conditional expectation of the potential outcomes, $E[Y|A = a, X] \neq E[Y(a)|X]$, preventing the identification of the optimal ITR. However, when a suitable instrument is available, one can still identify an optimal treatment rule in the presence of unmeasured confounding.

An instrumental variable Z is a variable related to the outcome Y only through the treatment A that allows for the identification of unbiased aspects of the treatment effect of A on Y, when the treatment is thought to be confounded by an unmeasured variable U. A valid IV relies on three core assumptions (see Figure 3.1), which we write in two

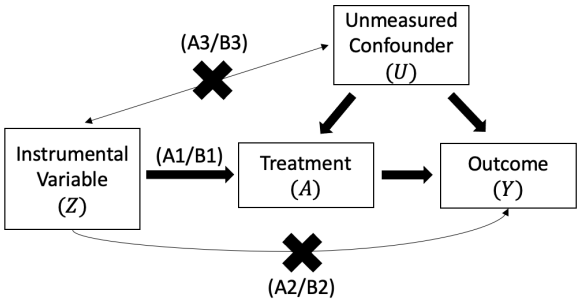


Figure 3.1: Directed acyclic graph representing the core assumptions for a valid instrument.

ways. The first assumption is IV relevance, that the instrument is related to the treatment, (A1) $E[\mathbb{1}\{A(z) = 1\}|X = x] - E[\mathbb{1}\{A(z') = 1\}|X = x] \neq 0$ for all x and $z \neq z'$ and (B1) $Z \not\perp A|X$. Here, we make the distinction between (A1) and (B1) so that under (A1), the IV has a causal relationship with the treatment and under (B1) the instrument need only be correlated. The second assumption is the exclusion restriction, that the instrument is

not directly related to the outcome, (A2/B2) $Y_i(z,a) = Y_i(z',a) = Y_i(a)$ for $z \neq z'$. The third assumption is IV ignorability, that the instrument is unrelated to any unmeasured confounders of the treatment's effect on the outcome, (A3) $A(z), Y\big(z, A(z)\big) \perp Z|X$ for all z and $A(z)$, and (B3) $Z \perp U|X$ and $Y(z,a) \perp (Z,A)|X,U$ for all z and a. The distinction between (A3) and (B3) is that (A3) requires any confounder of the instrument and potential treatments has been measured, and (B3) allows for an unmeasured confounder but requires the IV to be conditionally independent of the unmeasured confounder. The three assumptions IV relevance, exclusion restriction, and IV ignorability are necessary for an IV to identify unbiased bounds of the treatment effect [Balke and Pearl (1997)]. To point identify a treatment effect, an additional assumption is needed, namely monotonicity [Angrist et al. (1996)], or that there is no additive U-Z interaction in $E[A|X,U,Z]$, [Wang and Tchetgen Tchetgen (2018)]. The assumption of monotonicity, with (A1)-(A3), allows for point identification of the treatment effect of the complier subpopulation, and the assumption of no additive U-Z interaction in $E[A|X,U,Z]$, with (B1)-(B3), allows for point identification of the treatment effect of the overall population. We derive methods to identify optimal ITR using either point identification assumption with multilevel IV, as these two assumptions lead to different interpretations of treatment effects and different interpretations of ITR.

Cui and Tchetgen Tchetgen (2021b) detail how to identify an ITR when unmeasured confounding on the treatment effect is present using either the monotonicity or the no additive U-Z interaction point identification assumptions, when a valid binary instrument is available. However, many instruments commonly used in health studies are not actually binary instruments, such as distance to the specialty care provider [Mcclellan et al. (1994); Lorch et al. (2012)], physician or hospital preference [Brookhart and Schneeweiss (2007); O'Malley et al. (2011); Columbo et al. (2018); Huling et al. (2019); Martínez-Camblor et al. (2019a,b)], multiple genetic variants [Kang et al. (2016b); Wang and Kang (2021); Ye et al. (2021)], timing of admission [Malkin et al. (2000); Goyal et al. (2013)], and insurance plan [Cole et al. (2006)]. While it is possible to dichotomize the non-binary instruments, doing so may result in a loss of information and worsen estimation, resulting in a suboptimal ITR. For example, consider a multilevel IV with $k = 4$ levels where, for some $X = x$, the first 3 levels result in the same weak push toward taking the treatment, $P(A = 1|X = x, Z = z) = 0.25$ for $z = 1, 2, 3$, and the 4th level of the IV results in a stronger push, $P(A = 1|X = x, Z = z) = 0.5$ for $z = 4$. If an investigator was unaware that the highest level of the IV resulted in the stronger push toward taking the treatment, they

may dichotomize the multilevel IV at the median, resulting in the effect of the instrument on the IV as $P(A = 1|X = x, Z = 1) - P(A = 1|X = x, Z = 0) = 0.125$ and artificially creating a weak instrument. Further, the challenge of adequately dichotomizing a non-binary IV grows as the number of levels k of the multilevel IV grows. To better retain and utilize the information in the non-binary IV, we propose a generalization of the work in Cui and Tchetgen Tchetgen (2021b) using monotonicity and no additive U-Z interaction point identification assumptions along with contrast statements. Our work can be seen to be a generalization as it reduces to the methods identifying an ITR with binary instruments when $k = 2$.

To generalize to multilevel instruments, we use contrasts, defined as a linear combination of variables whose coefficients sum to zero $\sum_{z=1}^{k} c_z = 0$, to provide a general framework under both sets of assumptions, (A) and (B), to identify the treatment effects, corresponding value functions, and improve estimation. Without loss of generality, as the coefficients can be standardized, we also require the coefficients to have support $c_z \in [-1, 1]$. This results in many possible contrasts to combine and compare the different levels of the IV. For example, the difference of any two levels of the multilevel instrument, z and z', can be taken by setting level z to have coefficient $c_z = -1$, level z' to have coefficient $c_{z'} = 1$, and the rest of the coefficients to be 0. Other possibilities include using polynomial contrasts, such as, for $k = 3$ levels, the linear contrast $c_1 = -1$, $c_2 = 0$, and $c_3 = 1$ if the investigator believes the relationship between the IV and treatment is linear, or the quadratic contrast $c_1 = 0.5$, $c_2 = -1$, and $c_3 = 0.5$ if the investigator believes the relationship between the IV and treatment is quadratic. The choice of contrast is a reflection of the investigator's expectation of the instrument's relationship with the treatment and is largely made to combine and compare levels of the instrument so that the contrast of the probability of receiving the treatment is forced away from 0. For binary IV, there is only one contrast available, which is the linear contrast taking the difference between the two instruments. By dichotimizing a multilevel IV, this single contrast can result in different levels of the IV, that have different effects on the treatment reception, being grouped together, possibly resulting in poor estimation of treatment effects and a suboptimal ITR.

In the following two sections, we detail how to identify an ITR with a multilevel IV, what value function is being optimized, and how to interpret the identified ITR for the two target populations with their respective sets of assumptions.

3.2.3 Identifying ITR with Multilevel IV for the Complier Subpopulation

To identify optimal ITR with multilevel IV for the complier subpopulation, we assume the core IV assumptions (A1)-(A3) and the point identification assumption (A4) monotonicity, which states that no individual acts against the increasing levels of the IV. That is, there is no individual i such that $A_i(z) > A_i(z')$ for any $z < z'$ [Imbens and Angrist (1994)]. Without loss of generality, for monotonicity, we're assuming that an increase in the instrument level leads to an increase in the treatment. We note that this assumption imposes a sort of ordering on the support of the instrument, such that if patients are more likely to take the treatment given the instrument value $Z = z'$ than given $Z = z$, then any patient who would take the treatment for the instrument value $Z = z$ must also take the treatment for the value $Z = z' > z$.

Motivated by Angrist et al. (1996), Frangakis and Rubin (2002), and Cheng and Small (2006), we stratify individuals by their post-randomization potential treatment values. With a multilevel instrument, there are different kinds of compliers, which we refer to as sub-compliers. These sub-compliers can be defined by the level of the instrument for which their treatment behavior changes. That is, an individual i is an ℓ-complier if the level $\ell_i \in \{2, \ldots, k\}$ of the IV changes the ith individual's potential treatment from $A_i(1) = \cdots = A_i(\ell - 1) = -1$ to $A_i(\ell) = \cdots = A_i(k) = 1$. Under monotonicity, the ℓ-compliers can be defined as $A(\ell - 1) < A(\ell)$. For example, in a randomized encouragement design with varying degrees of encouragement, where the IV is the encouragement, an ℓ-complier is an individual who opts to take the treatment $A = 1$ if they receive at least the level ℓ of encouragement, but would otherwise take the treatment $A = -1$, if they receive any lower level of encouragement. We can then define the subpopulation of all ℓ-compliers as the union of all $k - 1$ sub-compliers, or succinctly, the individuals such that $A(1) < A(k)$ under monotonicity. We further define the usual always-takers, never-takers, and defiers. Always-takers always take the treatment regardless of their instrument values, $A(z) = 1$ for all $Z = z$. Never-takers never take the treatment regardless of their instrument values, $A(z) = -1$ for all $Z = z$. Defiers act against their instrument values, $A(z) > A(z')$ for any $z < z'$. Monotonicity then states that no defiers exist.

We define the conditional treatment effects of each sub-complier, and the unions of different sub-compliers, by using the contrasts defined previously. We note that these contrasts are not necessary for the definition of these local causal effects but rather a

convenient framework that can generally express any local causal effect of interest. Under assumptions (A1)-(A4), the ℓ-complier conditional average treatment effect is defined as $E[Y(1) - Y(-1)|X, A(\ell-1) < A(\ell)]$, which we note can be equivalently written as $E[Y(1) - Y(-1)|X, \mathbb{1}\{A(\ell) = 1\} - \mathbb{1}\{A(\ell-1) = 1\} = 1]$. Using our contrasts, we can simply write this conditional local effect as

$$E\left[Y(1) - Y(-1)\middle| X, \sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1\right], \tag{3.1}$$

where the contrast coefficients used are defined as $c_{\ell-1} = -1$, $c_\ell = 1$, and $c_z = 0$ for $z \neq \ell$. We denote these contrast coefficients as contrast $c_z^{(\ell)}$. Under monotonicity, the contrast of the indicators of the potential treatments, $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\}$, is equal to 1 for ℓ-compliers when using the contrast $c_z^{(\ell)}$. Also, the contrast of the indicators of the potential treatments using $c_z^{(\ell)}$ sums to 0 for all other possible principal strata, as allowed by montonicity. Therefore, (3.1) defines the local effects of the ℓ-complier when using the contrasts $c_z^{(\ell)}$. Further, we can use (3.1) to define the local effects of the unions of different sub-compliers, where one only needs to use the appropriate contrast coefficients. These contrast coefficients are derived by taking the sum across the $c_z^{(\ell)}$ contrasts of the ℓ-compliers contained in the union of interest. In other words, as the contrasts can be expressed as a vector in $[-1, 1]^k$, the contrast coefficients for the union of ℓ-compliers is the vector that is equal to the sum of the $c_z^{(\ell)}$ vectors of the individual ℓ-compliers contained in the union. For example, let $k = 5$ and let the effects of the union of the 3- and 5-compliers be of interest, $E[Y(1) - Y(-1)|X, A(2) < A(3)$ or $A(4) < A(5)]$. The contrasts of the 3- and 5-compliers can be written as the vectors $c_z^{(3)} = (0, -1, 1, 0, 0)$ and $c_z^{(5)} = (0, 0, 0, -1, 1)$, respectively. The contrast for the union of the 3- and 5-compliers is then $c_z^{(3,5)} = c_z^{(3)} + c_z^{(5)} = (0, -1, 1, -1, 1)$ and can be used with (3.1) to define their local effects. Under monotonicity, using the the contrast $c_z^{(3,5)}$, the contrast of the potential treatment indicators is equal to 1 only for both the 3- and 5-compliers, and equal to 0 for all other principal strata (always-takers, never-takers, 2-compliers, and 4-compliers). Therefore, (3.1) with the contrast $c_z^{(3,5)}$ defines the local effects of the union of the 3- and 5-compliers. Finally, the local effects for the union of all ℓ-compliers, $E[Y(1) - Y(-1)|X, A(1) < A(k)]$, can be seen to be defined by (3.1) with the contrast $c_z^{(all)}$ which is defined as $c_1 = -1$, $c_k = 1$, and $c_z = 0$ for $z \in \{2, \ldots, k-1\}$. That is, the contrast taking the difference between the highest level and lowest level of the IV defines the local effects for the union of all

ℓ-compliers.

The same contrasts used to define the ℓ-complier local effects can be used to identify their local effects. Following a similar argument presented in Angrist et al. (1996), under assumptions (A1)-(A4), the conditional local effects of the ℓ-complier or a union of ℓ-compliers is identified by what we refer to as the contrast-specific Wald estimand,

$$\frac{\sum_{z=1}^{k} c_z E[Y|X, Z=z]}{\sum_{z=1}^{k} c_z P(A=1|X, Z=z)} = E\left[Y(1)-Y(-1)\middle| X, \sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right], \tag{3.2}$$

where the contrasts used are those that correspond to the sub-compliers of interest. For example, using (3.2), the contrast $c_z^{(all)}$ identifies the local effects for the union of all ℓ-compliers, and contrast $c_z^{(\ell)}$ identifies the local effects for the ℓ-compliers. The contrast-specific Wald estimand is similar to the usual Wald estimand [Wald (1940)] in the binary IV setting, where the ratio of the intention-to-treat (ITT) is divided by the effect of the instrument on receiving the treatment. The difference in the multilevel IV setting from the binary IV setting is in the use of a contrast, as it is still a ratio of an ITT effect and the effect of the instrument on the treatment. Similar to the binary IV setting, the denominator of the contrast-specific Wald estimand can, under monotonicity and with the appropriate contrast, be interpreted as the probability of being an ℓ-complier or the probability of belonging to a union of ℓ-compliers. For example, using the contrast $c_z^{(\ell)}$, the denominator of the contrast-specific Wald estimand is the probability of being an ℓ-complier, and the contrast $c_z^{(all)}$ leads to the probability of being any of the $k-1$ ℓ-compliers. Now that we can identify the complier and ℓ-complier conditional treatment effects, we can determine how to identify an optimal ITR.

The optimal ITR for the ℓ-compliers, or union of ℓ-compliers, is defined using their respective conditional local treatment effects. This is written as,

$$\begin{aligned} d^*_{(A)}(X) &= \text{sign}\left\{ E\left[Y(1)-Y(-1)\middle| X, \sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right]\right\} \\ &= \arg\max_{d\in\mathcal{D}} E\left[Y(d(X))\middle| \sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right], \end{aligned}$$

where the contrasts used correspond to the ℓ-compliers, or union of ℓ-compliers, that are of interest. The choice of (A) in the subscript is used to denote the set of assumptions (A)

under which this optimal ITR is defined. The value functions this decision rule maximizes are defined as the expected outcome among the sub-compliers of interest for a given decision. That is, the value function for the ℓ-compliers, or union of ℓ-compliers, can be generally defined as,

$$V_{(A)}(d(X)) = E\left[Y(d(X))\middle|\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right],$$

where the contrasts used correspond to the ℓ-compliers, or union of ℓ-compliers, that are of interest. Similar to the identification of an ITR without unmeasured confounding, an ITR for the ℓ-compliers, or a union of ℓ-compliers, can be identified by using their respective estimated treatment effects. That is, under the assumptions (A1)-(A4), an ITR for the ℓ-compliers, or a union of ℓ-compliers, can be identified by taking the sign of the corresponding contrast-specific Wald estimand. However, as similarly shown in Cui and Tchetgen Tchetgen (2021b) and Qiu et al. (2021) for binary instruments, the ITR for ℓ-compliers, or a union of ℓ-compliers, can be identified by their corresponding ITT contrasts alone. That is, the denominator of the contrast-specific Wald estimand is not needed. This brings us to our first result using the set of assumptions (A). The proof can be found in Appendix B.1.

Theorem 3.1. *Under the assumptions (A1)-(A4), the optimal ITR for ℓ-compliers, or a union of ℓ-compliers, is identified by*

$$\begin{aligned} d^*_{(A)}(X) &= \text{sign}\left\{E\left[Y(1)-Y(-1)\middle|X,\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right]\right\} \\ &= \text{sign}\left\{\sum_{z=1}^{k} c_z E[Y|X,Z=z]\right\}, \end{aligned}$$

where the contrast $c_z^{(\ell)}$ corresponds to the optimal ITR for the ℓ-compliers, and the contrast for a union of ℓ-compliers can be derived by summing across the ℓ-complier contrasts.

Theorem 3.1 states that, for multilevel IV, the sign of the ITT contrast is equivalent to the sign of the ℓ-complier, or union of ℓ-compliers, local treatment effects, under the appropriate contrast. This can be a counter-intuitive result, as it implies that the optimal ITR can be identified for ℓ-compliers without ever observing what treatment the individuals received. Though surprising, some intuition for this fact may be gained by considering the outcomes of the different possible subpopulations, always-takers, never-takers, and

ℓ-compliers. For always-takers ($A(z) = 1$ for all $Z = z$), the outcome is always $Y = Y(1)$, and similarly, for never-takers ($A(z) = -1$ for all $Z = z$), the outcome is always $Y = Y(-1)$. Under monotonicity there are no defiers, so for ℓ-compliers we have $A(\ell - 1) < A(\ell)$, and thus, the outcome is $Y = \left(\sum_{z=1}^{\ell-1} \mathbb{1}\{Z = z\}\right) Y(-1) + \left(\sum_{z=\ell}^{k} \mathbb{1}\{Z = z\}\right) Y(1)$, a function of only Z. Therefore, under monotonicity, the exclusion restriction, and SUTVA, the IV alone is sufficient for observing the potential outcome for compliers and ℓ-compliers.

Theorem 3.1 allows for the ITR estimation by indirect methods, where an outcome model is used to identify the ITR. It is possible to also identify the optimal treatment rule for ℓ-compliers, or a union of ℓ-compliers, using the direct method outcome weighted learning with a multilevel IV. Using the respective contrasts for ℓ-compliers, or a union of ℓ-compliers, under the assumptions (A1)-(A4), a rule $d^* \in \mathcal{D}$ that maximizes the value function $V_{(A)}(d)$ is one that maximizes

$$\begin{aligned}\arg\max_{d\in\mathcal{D}} V_{(A)}(d) &= \arg\max_{d\in\mathcal{D}} E\left[Y(d)\middle|\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1\right] \\ &= \arg\max_{d\in\mathcal{D}} E\left[\sum_{z=1}^{k} \frac{c_z \mathbb{1}\{Z = z\} Y \mathbb{1}\{d(X) = A\}}{P(Z = z|X)}\right].\end{aligned}$$

This is a generalization of Theorem 6.2 in Cui and Tchetgen Tchetgen (2021b) to multilevel IV and follows from a similar argument and taking note that $\sum_{z=1}^{k} c_z = 0$. Though the above OWL expression is written with $\mathbb{1}\{d(X) = A\}$, it does not in fact require the treatment A to be measured, which coincides with the conclusion from Theorem 3.1. For ℓ-compliers, $A = -1$ when $Z = \ell - 1$ and $A = 1$ when $Z = \ell$, so the treatment is known given the instrument. Therefore, the optimal decision rule for ℓ-compliers, or a union of ℓ-compliers, can be identified via OWL using only the instrument.

While using either the sign of the ITT contrast or the OWL expression to identify the optimal ITR for ℓ-compliers is valid under the assumptions (A1)-(A4), it is challenging to apply such rules when we cannot be certain for which sub-complier type an individual belongs. However, an investigator may predict which sub-compliance type a patient belongs by estimating the denominator of the contrast-specific Wald estimand,

$$\sum_{z=1}^{k} c_z P(A = 1|X, Z = z) = P\left(\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1\middle|X, Z = z\right),$$

using the contrasts $c_z^{(\ell)}$. Because $\sum_{z=1}^{k} c_z^{(\ell)} \mathbb{1}\{A(z) = 1\} = 1$ only for ℓ-compliers, $\sum_{z=1}^{k} c_z^{(\ell)} P(A = 1|X, Z = z)$ can be used to predict whether or not an individual is an ℓ-complier. This then allows for the assignment of the optimal rule for ℓ-compliers to the individuals predicted as ℓ-compliers. Concisely, predicting the sub-compliance type and applying the corresponding sub-complier rule can be written as a single rule as follows

$$\begin{aligned} d(X, \hat{L}) &= \text{sign}\left\{ \sum_{\ell=2}^{k} \sum_{z=1}^{k} c_z^{(\ell)} P(A = 1|X, Z = z) d^*_{(\ell)}(X) \right\} \\ &= \text{sign}\left\{ \sum_{\ell=2}^{k} P(A(\ell) > A(\ell-1)|X) d^*_{(\ell)}(X) \right\}, \end{aligned}$$

where $d^*_{(\ell)}(X) = d^*_{(A)}(X)$ when using the contrast coefficients $c_z^{(\ell)}$ and is the optimal ITR for the ℓ-compliers. The rule $d(X, \hat{L})$ can be interpreted as a weighted vote for which treatment the individual should receive with weights equal to the probability of being an ℓ-complier. The treatment with plurality is then assigned for the given X. As the notation $d(X, \hat{L})$ implies, the decision is now a function of the observed covariates X and the predicted sub-compliance type $\hat{L}$. Further, $d(X, \hat{L})$ no longer aims to optimize the value function for the ℓ-compliers given the observed covariates X ($V_{(A)}(d(X))$ with the contrast $c_z^{(\ell)}$), but rather the value function for all ℓ-compliers given the observed X and unobserved sub-compliance type L, $V(d(X, L)) = E[Y(d(X, L))|A(1) < A(k)]$. The value function $V(d(X, L))$ can be interpreted as the expected outcome of the whole complier subpopulation, or the union of all ℓ-compliers, for a given decision that depends on both the observed effect modifiers X and latent effect modifier L. Due to the necessary step of predicting $\hat{L}$, an investigator may prefer a different rule to $d(X, \hat{L})$, where the magnitude of the sub-compliers' effect is included in the weight, so that individuals who would benefit the most are weighted more heavily. For this reason, we propose the rule

$$\begin{aligned} \tilde{d}(X, \hat{L}) &= \text{sign}\left\{ \sum_{\ell=2}^{k} \sum_{z=1}^{k} c_z^{(\ell)} P(A = 1|X, Z = z) E[Y|X, Z = z] \right\} \\ &= \text{sign}\left\{ \sum_{\ell=2}^{k} P(A(\ell) > A(\ell-1)|X) \big(E[Y|X, Z = \ell] - E[Y|X, Z = \ell - 1] \big) \right\}. \end{aligned} \tag{3.3}$$

The rule $\tilde{d}(X, \hat{L})$ differs from $d(X, \hat{L})$ by replacing $d^*_{(\ell)}(X)$ with the ITT effect of the ℓ-

compliers. This change effectively lessens the harm from assigning $d^*_{(\ell)}(X)$ to an ℓ'-complier when the ℓ-complier and ℓ'-complier rules differ for a given X, $d^*_{(\ell)}(X) \neq d^*_{(\ell')}(X)$, for $\ell \neq \ell'$. That is, by including the ITT effects, we protect ourselves from assigning the wrong treatment to the compliers who would most benefit. The rule $\tilde{d}(X, \hat{L})$ can also be seen to be similar to the rule for the union of all ℓ-compliers, $d^*_{(all)}(X) = d^*_{(A)}(X)$ when using the contrast coefficients $c_z^{(all)}$. This rule is identified by taking the ITT difference between the highest and lowest level of the instrument and is equivalent to the sum of all ℓ-complier ITT effects, $d^*_{(all)}(X) = E[Y|X, Z = k] - E[Y|X, Z = 1] = \sum_{\ell=2}^{k} E[Y|X, Z = \ell] - E[Y|X, Z = \ell - 1]$. Both $\tilde{d}(X, \hat{L})$ and $d^*_{(all)}(X)$ assign treatment based on which ℓ-complier subpopulation has the larger ITT effect, but $\tilde{d}(X, \hat{L})$ also takes into account the probability that an individual belongs to the ℓ-complier subpopulation. The rule $\tilde{d}(X, \hat{L})$ can therefore better accommodate the setting where different ℓ-complier and ℓ'-complier rules differ for a given X than the rule $d^*_{(all)}(X)$. However, unless we can predict the sub-compliance type with certainty, $P(A(\ell) > A(\ell - 1)|X) = 1$ for all ℓ, so that the ITT effect of the ℓ-complier is used to assign treatment for the ℓ-complier, the rule $\tilde{d}(X, \hat{L})$ cannot be guaranteed to be optimal. That is, $\tilde{d}(X, \hat{L})$ cannot be guaranteed to maximize the value function $V(d(X, L))$, as shown in our second result,

Proposition 3.1. *Under the assumptions (A1)-(A4), we have that the decision maximizing* $V(d(X, L))$ *maximizes*

$$\begin{aligned}\arg\max_{d\in\mathcal{D}} V(d(X,L)) &= \arg\max_{d\in\mathcal{D}} E\left[Y\big(d(X,L)\big)\big|A(1) < A(k)\right] \\ &= \arg\max_{d\in\mathcal{D}} E\left[\sum_{\ell=2}^{k} \mathbb{1}\{d(X,\ell) = 1\}\big(E[Y|X, Z = \ell] - E[Y|X, Z = \ell - 1]\big)\right].\end{aligned}$$

Therefore, if $P(A(\ell) > A(\ell-1)|X) = 1$*, for all* ℓ*, we have that*

$$\begin{aligned} d^*(X,L) &= \arg\max_{d\in\mathcal{D}} E\left[\sum_{\ell=2}^{k} \mathbb{1}\{d(X,\ell)=1\}\big(E[Y|X,Z=\ell] - E[Y|X,Z=\ell-1]\big)\right] \\ &= \arg\max_{d\in\mathcal{D}} E\Bigg[\sum_{\ell=2}^{k} \mathbb{1}\{d(X,\ell)=1\}P(A(\ell) > A(\ell-1)|X) \\ &\qquad\qquad \times \big(E[Y|X,Z=\ell] - E[Y|X,Z=\ell-1]\big)\Bigg] \\ &= \text{sign}\left\{\sum_{\ell=2}^{k}\sum_{z=1}^{k} c_z^{(\ell)} P(A=1|X,Z=z)E[Y|X,Z=z]\right\}. \end{aligned}$$

Proposition 3.1 states that if one correctly predicted the sub-compliance type L with certainty, the decision rule $\tilde{d}(X,\hat{L})$ is optimal for the value function for compliers given the observed effect modifier X and latent effect modifier L, $V(d(X,L)) = E[Y(d(X,L))|A(1) < A(k)]$. The proof can be found in Appendix B.2.Though in practice, we cannot predict the sub-compliance type L with certainty, a benefit to using $\tilde{d}(X,\hat{L})$ is in reducing potential harm from using the rule for all ℓ-compliers, $d^*_{(all)}(X)$. If different ℓ-compliers, ℓ and ℓ', have treatment effects with opposite sign, using the rule $d^*_{(all)}(X)$ will cause harm to either the ℓ-compliers or $\ell'-$compliers, whichever have a lower ITT effect, and therefore a lower value. Alternatively, $\tilde{d}(X,\hat{L})$ takes into account the probability of being an ℓ-complier and an aspect of their treatment effect in order to identify which ℓ-complier is most likely and whether they most benefit to then assign the corresponding decision.

In order to evaluate the performance of a given decision rule, we must identify the value function being optimized. To estimate the value function $V_{(A)}(d)$ for a given decision rule d, we can leverage an IPW estimator, as in outcome weighted learning. That is, under the assumptions (A1)-(A4), we can write $V_{(A)}(d)$ in terms of observables,

$$V_{(A)}(d) = E\left[Y(d)\middle|\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\} = 1\right] = E\left[\sum_{z=1}^{k} \frac{c_z \mathbb{1}\{Z=z\}AY\mathbb{1}\{d(X)=A\}}{\delta_c P(Z=z|X)}\right], \tag{3.4}$$

where $\delta_c = \sum_{z=1}^{k} c_z P(A=1|Z=z)$ and the contrast to be used corresponds to the ℓ-complier, or union of ℓ-complier, subpopulation of interest. This is a generalization of Theorem 6.1 in Cui and Tchetgen Tchetgen (2021b) to multilevel IV and follows

from a similar argument and taking note that $\sum_{z=1}^{k} c_z = 0$. Therefore, to estimate the value function $V_{(A)}(d)$, the empirical form of (3.4) can be used for a given decision and choice of contrast. However, the instrument propensity may be unknown for observational instruments satisfying monotonicity, so we additionally provide an augmented inverse probability weighted (AIPW) estimator to identify the complier and ℓ-complier value functions,

$$V_{(A)}(d) = E\left[\sum_{z=1}^{k} c_z \left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{P(Z=z|X)} + h_{az}(X)\right)\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c}\right], \tag{3.5}$$

where $h_{az}(X) = E[\mathbb{1}\{A=a\}Y|X, Z=z]$. Under the assumptions (A1)-(A4), this AIPW estimator identifies the value function $V_{(A)}(d)$ for a given decision and contrast that corresponds to the ℓ-complier, or union of ℓ-complier, subpopulation of interest. The proof can be found in Appendix B.3. The AIPW estimator, (3.5), is doubly robust in the sense that either the instrument propensity or the model for the interaction of the treatment and outcome, $h_{az}(X)$, need to be correctly specified, but not both, in order to identify the value function. We note that δ_c must also be correctly specified, regardless of whether or not the propensity score or $h_{az}(X)$ are correctly specified. For this reason, we suggest using the doubly robust estimator

$$\delta_c = E\left[\sum_{z=1}^{k} c_z \left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\} - P(A=a|X,Z=z)\big)}{P(Z=z|X)} + P(A=a|X,Z=z)\right)\right],$$

to ensure δ_c is correctly specified. The empirical form of the AIPW estimator, (3.5), can be used to estimate the value function $V_{(A)}(d)$ for a given decision and contrast that corresponds to the ℓ-complier, or union of ℓ-compliers, with either the instrument propensity $P(Z=z|X)$ or model $h_{az}(X)$ correctly specified.

3.2.4 Identifying ITR with Multilevel IV for the Overall Population

To identify optimal ITR with multilevel IV for the overall population, we assume the core IV assumptions (B1)-(B3) and the point identification assumption (B4) no additive U-Z interaction in $E[Y|X,U,Z]$, which states that if an unmeasured variable U confounds the treatment effect A on Y, it does not interact with the instrument Z on the additive scale in predicting the treatment. Succinctly, the assumption of no additive U-Z interaction in

$E[Y|X,U,Z]$ can be written as $\delta_c(X) = \tilde{\delta}_c(X,U)$, where $\delta_c(X) = \sum_{z=1}^{k} c_z P(A=1|X,Z=z)$ and $\tilde{\delta}_c(X,U) = \sum_{z=1}^{k} c_z P(A=1|X,U,Z=z)$. An implication of the notation for this assumption is that it depends on a choice of contrast. A plausible advantage to using a multilevel instrument and contrasts is that it allows for some levels of the IV to dissatisfy (B4), so long as we define contrast coefficients $c_z = 0$ for levels $Z = z$ of the IV that interact with U, and the other levels of the instrument do not additively interact with U. However, we cannot know which contrasts in fact satisfy (B4), so we assume that the assumption of no additive U-Z interaction holds for all contrasts.

Under the point identification assumption of no additive U-Z interaction, the overall population's average treatment effect is identifiable. Therefore, the value function an ITR optimizes in the setting with multilevel IV and assumptions (B1)-(B4) is the same as the value function optimized in the setting that assumes no unmeasured confounding, $V_{(B)}(d(X)) = E[Y(d(X))]$, the overall population outcome for a given decision. Here the subscript of (B) is used to denote the use of the set of assumptions (B) used to identify an optimal ITR. When using a binary IV, the overall population's average treatment effect is identified by the Wald estimand [Wang and Tchetgen Tchetgen (2018)]. Extending to a multilevel IV, we use the contrast-specific Wald estimand, (3.2), to identify the population's average treatment effect. Though, similar to Section 3.2.3 in the use of contrasts, there are two key differences. The first difference is that, without the (A4) monotonicity assumption, the contrasts can no longer identify ℓ-complier subpopulations. So, while the same contrasts may be used, they do not have the interpretation as in Section 3.2.3. The second difference is that the choice of contrast is not used to identify certain subpopulations, but rather an identifiability tool for the overall population's conditional average treatment effect. That is, under the assumptions (B1)-(B4), the contrast-specific Wald estimand is equal to the conditional average treatment effect,

$$\begin{aligned}
\frac{\sum_{z=1}^{k} c_z E[Y|X,Z=z]}{\sum_{z=1}^{k} c_z P(A=1|X,Z=z)} &= \frac{E\left[E\big\{Y(1)-Y(-1)\big|X,U\big\}\sum_{z=1}^{k} c_z P(A=1|X,U,Z=z)\Big|X\right]}{\sum_{z=1}^{k} c_z P\left(A=1|X,Z=z\right)} \\
&= E\left[\frac{E\big\{Y(1)-Y(-1)\big|X,U\big\}\sum_{z=1}^{k} c_z P(A=1|X,Z=z)}{\sum_{z=1}^{k} c_z P\left(A=1|X,Z=z\right)}\middle|X\right] \\
&= E\left[Y(1)-Y(-1)|X\right],
\end{aligned}$$

where the point identification assumption (B4) is used in the second equality. Additional

details for the proof can be found in Appendix B.4. Therefore, any contrast can be used to identify the conditional average treatment effect, so long as the contrast satisfies the point identification assumption (B4) of no additive U-Z interaction.

While the no additive U-Z interaction assumption is necessary to identify the value function $V_{(B)}(d(X)) = E[Y(d(X))]$, it can be relaxed to identify the optimal ITR using the contrast-specific Wald estimand. For a binary IV, Cui and Tchetgen Tchetgen (2020) and Cui and Tchetgen Tchetgen (2021a) provide a necessary and sufficient condition to identify an optimal ITR under the core IV assumptions (B1)-(B3). We extend their necessary and sufficient condition for binary IV to multilevel IV for an arbitrary contrast. The proof can be found in Appendix B.4.

Theorem 3.2. *Under the set of assumptions (B1)-(B3), the following condition is necessary and sufficient for the sign of the contrast-specific Wald estimand to identify an optimal ITR for the overall population,*

$$E\left[\frac{\tilde{\gamma}(X,U)}{\gamma(X)} \times \frac{\tilde{\delta}_c(X,U)}{\delta_c(X)}\middle| X\right] > 0, \tag{3.6}$$

where $\tilde{\gamma}(X,U) = E[Y(1)-Y(-1)|X,U]$, $\gamma(X) = E[Y(1)-Y(-1)|X]$, $\tilde{\delta}_c(X,U) = \sum_{z=1}^{k} c_z P(A = 1|X,U,Z=z)$, *and* $\delta_c(X) = \sum_{z=1}^{k} c_z P(A=1|X,Z=z)$. *The optimal ITR for the overall population is then*

$$d^*_{(B)}(X) = \text{sign}\left\{E\left[Y(1) - Y(-1)|X\right]\right\} = \text{sign}\left\{\frac{\sum_{z=1}^{k} c_z E[Y|X,Z=z]}{\sum_{z=1}^{k} c_z P(A=1|X,Z=z)}\right\}.$$

The necessary and sufficient condition (3.6) for multilevel IV and an arbitrary contrast can be loosely interpreted as the four estimands, or pairs of the four estimands, $\tilde{\gamma}(X,U)$, $\gamma(X)$, $\tilde{\delta}_c(X,U)$, and $\delta_c(X)$, sharing the same sign on average, conditional on the observed X. What this assumption rules out is the setting where three of the estimands share the same sign, but one is different from the rest on average, conditional on X. For example, an optimal ITR is unattainable in the setting $\tilde{\gamma}(X,U) > 0$, $\gamma(X) > 0$, $\tilde{\delta}_c(X,U) > 0$, but $\delta_c(X) < 0$ for all X and U. The signs of the estimands are allowed to differ for some U, so long as, conditioning on X, on average they are the same. Cui and Tchetgen Tchetgen (2020) interpret this condition as allowing for the identification of the optimal ITR in the setting "where individuals' decision to uptake the intervention is concordant with an

anticipated benefit from the intervention", or where individuals opt into the treatment they believe will best benefit them. The generalization to multilevel IV, where a contrast is used, changes this interpretation only slightly on how individuals opt into the treatment, where the contrast is a statement on how individuals behave with the different levels of the IV.

Theorem 3.2 allows for the ITR estimation by indirect methods, where outcome models for the numerator and denominator of the contrast-specific Wald estimand is used to identify the ITR. The necessary and sufficient condition also allows for the direct method using OWL with a multilevel IV. That is, under the assumptions (B1)-(B3), a rule $d^* \in \mathcal{D}$ that maximizes the population value function is

$$\arg\max_{d\in\mathcal{D}} E\left[Y(d)\right] = \arg\max_{d\in\mathcal{D}} E\left[\sum_{z=1}^{k} \frac{c_z \mathbb{1}\{Z=z\}AY\mathbb{1}\{d(X)=A\}}{\delta_c(X)P(Z=z|X)}\right], \tag{3.7}$$

whenever the necessary and sufficient condition (3.6) holds. This is a generalization of Theorem 3.1 in Cui and Tchetgen Tchetgen (2020) to multilevel IV and follows from a similar argument and taking note that $\sum_{z=1}^{k} c_z = 0$.

Though an optimal ITR is identifiable using either the contrast-specific Wald-estimand or OWL whenever the necessary and sufficient condition holds, to estimate the ITR a contrast needs to be chosen that satisfies the condition. Since we cannot know which contrasts satisfy the necessary and sufficient condition, we assume that the condition is satisfied for any contrast. However, there is still a choice of contrast to be made, where the decision now is one of estimation performance. To avoid making a poor decision in contrast, one that negates the effect of the instrument on the probability of receiving the treatment, we propose learning the contrasts from the data. While it is possible for an investigator to try several different contrasts, seeing which contrasts do not artificially introduce a weak instrument bias by negating the effect of the instrument on receiving the treatment, this solution quickly becomes intractable as the number of contrasts to consider grows rapidly as the levels k of the IV increases. Additionally, for some values of the observed covariates X, one contrast may perform better than another. That is, when the instrument interacts with X in the contrast effect of the IV on the treatment, one contrast may negate the effect of the instrument on the probability of receiving the treatment while another does not, depending on the values of X. This would result in the investigator needing to try many different contrasts for the different values of X. Instead, we propose a contrast that forces the ITT contrast to better resemble the conditional average treatment effect. Such a

contrast may be of interest for identifying the optimal ITR because it does not require a point identification assumption and because the optimal treatment rule is equivalent to the sign of the conditional average treatment effect. Further, as we will show, such a contrast will prevent the use of a contrast that negates the effect of the instrument on the probability of receiving the treatment. Such a contrast can be written in the form,

$$M^*(X) = \arg\min_m \left| E\big[Y(1) - Y(-1)\big|X\big] - \sum_{z=1}^{k} c_z^{(m)} E[Y|X, Z=z] \right|,$$

where m indexes the possible contrasts. An implication of this expression is that the contrast forcing the ITT to resemble the conditional average treatment effect is one that is tailored to an individual; this idealized contrast is a function of the observed covariates X. In effect, the use of the this contrast $M^*(X)$ is to use different contrasts for different individuals to identify the optimal ITR, effectively individualizing ITR estimation.

Under the assumptions (B1)-(B3), the ITT contrast can be rewritten so that

$$\sum_{z=1}^{k} c_z E[Y|X, Z=z] = E\left[E\big[Y(1) - Y(-1)\big|X, U\big] \left(\sum_{z=1}^{k} c_z P(A=1|X, U, Z=z) \right) \middle| X \right].$$

This implies that the contrast $M^*(X)$ is equivalent to the contrast that forces $\sum_{z=1}^{k} c_z P(A = 1|X, U, Z)$ close to 1. Because the contrast coefficients have support $c_z \in [-1, 1]$ and are constrained so that $\sum_{z=1}^{k} c_z = 0$, the probability contrast $\sum_{z=1}^{k} c_z P(A = 1|X, U, Z)$ is bound between -1 and 1. Therefore, the contrast $M^*(X)$ is equivalent to

$$M^*(X) = \arg\max_m \sum_{z=1}^{k} c_z^{(m)} P(A=1|X, U, Z=z).$$

Therefore, the contrast $M^*(X)$ is effectively forcing the effect of the instrument on the treatment away from 0, ensuring a contrast that may inadvertently negate the effect of the instrument is not used. However, as $P(A = 1|X, U, Z = z)$ is in terms of the unmeasured U, we cannot directly estimate this probability. The following result shows that the contrast $M^*(X)$ is still identifiable. The proof can be found in Appendix B.5.

Proposition 3.2. *Define the contrast $M(X)$ as*

$$M(X) = \arg\max_m \left| \sum_{z=1}^{k} c_z^{(m)} E[Y|X, Z = z] \right|.$$

Under the assumptions (B1)-(B3), we have

$$M(X) = M^*(X).$$

The contrast $M(X)$ is the argmax of the absolute value of the ITT contrasts and is a function of the observed covariates X. Proposition 3.2 allows for the learning of contrast coefficients, effectively avoiding a single choice of contrast to identify an optimal ITR. Though no assumption or condition using a contrast is needed to derive $M(X)$, the contrasts $M(X)$ must satisfy the necessary and sufficient condition (3.6) in order for it to be used to identify an optimal ITR.

We could similarly use the contrast $M(X)$ under the set of assumptions (A) for the complier subpopulation, though it will always result in the same contrast for all X under the monotonicity assumption. That is, under the set of assumptions (A), the contrast $M(X)$ is no longer a function of X, but rather a constant contrast. To see this note that the ITT contrast can be rewritten, under the assumptions (A1)-(A4) as

$$\sum_{z=1}^{k} c_z E[Y|X, Z = z] = E\left[E\big(Y(1) - Y(-1)\big) \sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} \middle| X \right].$$

Therefore, for the set of assumptions (A), the argmax of the absolute value of the ITT contrast is maximizing $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} \in [-1, 1]$. Under monotonicity, so that no defiers exist, we have that $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 0$ always for both never-takers and always-takers, and for the contrast coefficients $c_z^{(all)}$, $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1$ for all ℓ-compliers. This results in the argmax contrast being equivalent to the single contrast that identifies the average treatment effect for all ℓ-compliers, $M(X) = c_z^{(all)}$. The structure that monotonicity imposes on the effect of the IV on the treatment provides a single solution for the contrast learned from the argmax of the ITT contrasts.

Under either settings of assumptions, (A) for the complier subpopulation or (B) for the overall population, are satisfied, the argmax contrast $M(X)$ can be used to identify the optimal ITR. The decision rule using the contrast $M(X)$, which we refer to as the argmax

rule, is written as

$$d_M(X) = \text{sign}\left\{ \frac{\sum_{z=1}^{k} c_z^{(M(X))} E[Y|X, Z=z]}{\sum_{z=1}^{k} c_z^{(M(X))} P(A=1|X, Z=z)} \right\}. \tag{3.8}$$

As shown in Section 3.2.3, the contrast-specific Wald estimand or the ITT contrast with contrast coefficients $c_z^{(all)}$ can be used to identify the optimal ITR for all of the compliers, so using $M(X) = c_z^{(all)}$ under monotonicity will identify the optimal ITR for all of the compliers. For the assumptions (B), we assume that the necessary and sufficient condition holds for all contrasts used in $M(X)$, therefore, by Theorem 3.2, the argmax decision $d_M(X)$ using the contrasts $M(X)$, can be used to identify the optimal ITR for the overall population, $d_M(X) = d^*_{(B)}(X)$.

To identify the value function for the overall population and evaluate the performance of a given decision rule, we leverage either an IPW estimator or an AIPW estimator. For the IPW estimator, assuming (B4) no additive U-Z interaction, we can use the empirical form of (3.7), as under assumption (B4) the value function, $V_{(B)}(d)$, is equivalent to the outcome weighted learning expression, (3.7). That is, the empirical form of outcome weighted learning expression estimates the value function $V_{(B)}(d)$ for the overall population when the assumption of no additive U-Z interaction is satisfied. In the case that the instrument propensity is unknown, we suggest an AIPW estimator to identify the overall population value function. Under the assumptions (B1)-(B4), the value function for the overall population $V_{(B)}(d) = E[Y(d)]$, is identified by

$$V_{(B)}(d) = E\left[\sum_{z=1}^{k} c_z \left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{P(Z=z|X)} + h_{az}(X) \right) \frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)} \right], \tag{3.9}$$

where $h_{az}(X)$ is defined as before for the AIPW estimator under monotonicity (3.5). The proof can be found in Appendix B.6. The AIPW estimator (3.9) is doubly robust in the sense that either the instrument propensity $P(Z=z|X)$ or the model $h_{az}(X)$ must be correctly specified, but not both, in order to identify the value function $V_{(B)}(d)$. Further, any contrast can be used to identify the value function, so long as it satisfies the assumption of no additive U-Z interaction. We note that, regardless of whether or not the instrument propensity or $h_{az}(X)$ is correctly specified, $\delta_c(X)$ must be correctly specified for the AIPW estimator (3.9) to identify the value function. The empirical form of this AIPW estimator

can then be used to estimate the overall population value function for contrasts that satisfy the assumption of no additive U-Z interaction when either the instrument propensity or model $h_{az}(X)$ is correctly specified.

As was the case in identifying the optimal ITR, to identify the value function $V_{(B)}(d)$ there is a choice of contrast to be made. Though we assume that the assumption of no additive U-Z interaction holds for all contrasts so that any choice can identify the value function, the choice of contrast can result in different performances in estimation. Namely, the choice of contrast should be made to mitigate the possibility of selecting a contrast such that $\sum_{z=1}^{k} P(A = 1|X, U, Z = z)$ is close to 0. To achieve this, we again use the contrasts $M(X)$ used for the argmax rule. That is, we use the learned contrasts $M(X)$ that maximize the absolute value of the ITT contrasts, which results in contrast coefficients forcing $\sum_{z=1}^{k} c_z P(A = 1|X, U, Z = z)$ toward 1. Using the learned contrasts removes the need to choose a single contrast and protects from inadvertently using contrasts that negate the effect of the instrument on receiving the treatment.

However, there may be concern that, by using the argmax contrasts $M(X)$, we use information on the outcome to evaluate the performance of the decision functions. It is possible that this could introduce bias in the estimation of the value function. Because $M(X)$ is defined as the absolute value of the ITT contrast, there will exist correlations between the chosen contrasts and the expected outcome. Though, this would affect the variance of the estimated value function, and therefore any inference, we do not expect it to be a cause of concern in our use. For sufficiently large samples, the empirical form of the AIPW estimator (3.9) or the IPW estimator (3.7) using the argmax contrasts should consistently estimate the value function. We demonstrate this numerically in Appendix B.8 using the AIPW estimator. In the event that the sample is not sufficiently large, we recommend choosing a contrast that does not negate the effect of the IV on the probability of receiving the treatment.

3.2.5 Bridging ITR Identification for the Two Sets of Assumptions

The two sets of assumptions (A1)-(A4) for the complier subpopulation and (B1)-(B4) for the overall population lead to different decision rules and value functions. The set of assumptions (A1)-(A4) targets the value function for ℓ-compliers through the use of specific contrasts $c_z^{(\ell)}$. In the event that the different ℓ-compliers have different decisions for a given X or when it is unclear which ℓ-complier rule to assign, it's required to predict the

latent sub-complier type and their respective decision rules to maximize the complier value function. The set of assumptions (B1)-(B4) targets the overall population value function and any contrast can be used as long as it satisfies the necessary and sufficient condition (3.6). However, a choice of contrast, or learning contrasts as a function of the observed covariates X, is still needed to maximize the overall population value function, where the contrast used has implications on estimation. Though the assumptions and their respective settings for identifying an optimal ITR are different, it is possible to bridge the two through the necessary and sufficient condition.

The necessary and sufficient condition (3.6) is a statement on the unmeasured variable U and its extent on modifying both the treatment effect and the propensity to take the treatment. For example, let $\tilde{\gamma}(X,U) \neq \gamma(X)$, but $sign\{\tilde{\gamma}(X,U)\} = sign\{\gamma(X)\}$, and similarly, $\tilde{\delta}_c(X,U) \neq \delta_c(X)$, but $sign\{\tilde{\delta}_c(X,U)\} = sign\{\delta_c(X)\}$ for all U. This setting can arise when U is an effect modifier that does not change the sign of the treatment effect or contrast of the probability of receiving the treatment (e.g. $\tilde{\gamma}(X,U) = \gamma(X)U^2$ and $\tilde{\delta}_c(X,U) = \delta_c(X)U^2$ for $U \in [-1,1]$). In such a setting, the variable U is an unmeasured effect modifier that modifies the effects of the treatment on the outcome and the instrument on the probability of receiving the treatment, but not in a way such that the condition (3.6) is violated. Further, in the event that $\tilde{\delta}_c(X,U)$ and $\delta_c(X)$ are both positive, the ITT contrast alone can identify the optimal ITR; the contrast-specific Wald estimand isn't needed. That is, the sign of the ITT contrast is the same as the sign of $\gamma(X)$. Though U is a latent effect modifier, it does not change the sign of the treatment effect nor the sign of the effect of the instrument on the treatment, and since the probability of receiving the treatment contrast is positive, the sign of the effect of the instrument on the outcome alone can identify the sign of the effect of the treatment on the outcome. So long as U modifies the effects of $\tilde{\gamma}(X,U)$ and $\tilde{\delta}_c(X,U)$ in a way that the necessary and sufficient condition is satisfied, then an optimal ITR is attainable using the sign of the contrast-specific Wald estimand or the outcome weighted learning expressions.

A similar necessary and sufficient condition with largely the same interpretation can be written for the setting using the assumptions (A1)-(A4) for the complier subpopulation. Under the assumptions (A1)-(A4), the necessary and sufficient condition for multilevel IV,

is written as

$$E\left[\frac{E[Y(1)-Y(-1)|X,L]}{\gamma(X)}\times\frac{P\left(\sum_{z=1}^{k}c_z\mathbb{1}\{A(z)=1\}=1\Big|X,A(1)<A(k)\right)}{\delta_c(X)}\Bigg|X\right]>0, \tag{3.10}$$

where L denotes the latent sub-compliance type, the contrasts considered are $c_z^{(\ell)}$, and $\gamma(X)$ and $\delta_c(X)$ are defined as above. The monotonicity assumption is used here to simplify this necessary and sufficient condition (3.10) to a similar form as the necessary and sufficient condition (3.6) for assumption setting (B); without the monotonicity assumption, the condition is slightly less succinct. The details and proof of the necessary and sufficient condition under assumptions (A1)-(A4) are in Appendix B.7. The necessary and sufficient conditions are clearly similar for the two sets of assumptions (A) and (B). The differences are in defining the unmeasured confounder as the latent sub-complier type L and conditioning on the union of all ℓ-compliers, but the interpretation is much the same for the necessary and sufficient condition for assumption setting (A).

The necessary and sufficient condition (3.10) is again a statement on latent variables, in this case L and $A(1)<A(k)$, and their extent on modifying both the treatment effect and the propensity to take the treatment. If the treatment effect for the ℓ-complier, $E[Y(1)-Y(-1)|X,L]$, and the probability of being an ℓ-complier given the individual is a a type of complier, $P\left(\sum_{z=1}^{k}c_z\mathbb{1}\{A(z)=1\}=1\Big|X,A(1)<A(k)\right)$, agree on average, conditional on X, in sign with $\gamma(X)$ and $\delta_c(X)$, then the sign of the contrast-specific Wald estimand using contrast coefficients $c_z^{(\ell)}$ identifying the ℓ-complier treatment effect will equal the sign of the overall treatment effect $\gamma(X)$. Here, the overall treatment effect is the average treatment effect for all compliers, never-takers, and always-takers. Because, under monotonicity, the $\delta_c(X)$ is the probability of being an ℓ-complier, and $P\left(\sum_{z=1}^{k}c_z\mathbb{1}\{A(z)=1\}=1\Big|X,A(1)<A(k)\right)$ is the probability of being an ℓ-complier given the individual is a type of complier, the right-hand side of the product in (3.10) will always be positive. That means, for the sign of the contrast-specific Wald estimand identifying the ℓ-complier treatment effect to identify the optimal ITR for the overall population, it must agree in sign with $\gamma(X)$. That is, the monotonicity assumption imposes a structure on the effect of the instrument on receiving the treatment, so that, to identify the optimal ITR for a population/subpopulation other than the ℓ-complier, the sign of the ℓ-complier treatment effect must agree in sign with the treatment effect of the other population/subpopulation of interest. Under the interpretation of the necessary

and sufficient conditions of how individuals opt into treatment, monotonicity imposes a structure so that ℓ-compliers must opt into treatment at the instrument level $Z = \ell$, so for the optimal ITR to be identified the ℓ-complier local effect must agree in sign with the treatment effect of the target population.

This interpretation coincides with the discussion in Section 3.2.3 on how the different ℓ-compliers may have different treatment rules for a given X, so that using the ℓ-complier rule, $d^*_{(\ell)}$, may be suboptimal for the union of all ℓ-compliers. The necessary and sufficient condition (3.10) states that, for the ℓ-complier treatment effect to identify the optimal ITR for all the ℓ-compliers, it must agree in sign with the treatment effect for all the compliers. This can be seen by removing $\gamma(X)$ (as we are now discussing the effect for only compliers) and rewriting the condition as

$$\frac{E[Y(1) - Y(-1)|X, L]}{E[Y(1) - Y(-1)|X, A(1) < A(k)]} > 0,$$

where the ratio of $P\left(\sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1 \middle| X, A(1) < A(k)\right)$ and $\delta_c(X)$ can be removed because they will always be positive under the monotonicity assumption. This also reflects the conclusion that the ITT contrast for the ℓ-complier can identify the optimal ITR for all the ℓ-compliers when the sign of the treatment effects of all the ℓ-compliers agree. In such a setting, the latent effect modifier L does not change the sign of the effect of the instrument on receiving the treatment due to the structure monotonicity imposes, nor does it change the sign of the treatment effect for the different ℓ-compliers. Therefore, the sign of the ITT contrast using contrast coefficients $c_z^{(\ell)}$ identifies the optimal ITR for the ℓ-compliers by Theorem 3.1, as well as the optimal ITR for all the ℓ-compliers when the latent sub-compliance type L does not change the signs of the different ℓ-complier treatment effects.

Under either set of assumptions (A) or (B), their respective necessary and sufficient conditions are satisfied when identifying the optimal ITR for the corresponding target populations. Due to the differences in assumption sets (A) and (B), the conditions are slightly different in form, but the interpretations are largely the same. For a population in which an unmeasured confounder, U or L, modifies the treatment effects, the unmeasured effect modifier must modify both the treatment effect and propensity to take the treatment so that the necessary and sufficient condition is satisfied in order to identify an optimal ITR.

3.3 Simulations

We conduct extensive simulation studies to measure the performance of the proposed decision rules, with a multilevel instrument with $k = 3$ levels, for the two sets of assumptions, (A1)-(A4) targeting the complier subpopulation, and (B1)-(B4) targeting the overall population. In all the simulation studies, we vary the instrument's association to the treatment because prior works have shown that performance of IV methods depends heavily on IV strength. In particular, problems can arise when the instrument has a weak effect on the treatment; see Staiger and Stock (1997), Stock, Wright, and Yogo (2002), and references therein for more details.

For comparisons, we estimate decision rules using a binary instrument, no instrument, and as an oracle observing the unmeasured variables. For the setting under assumptions (A) for the complier subpopulation, the binary instrument rule is the sign of the ITT effect, and for the setting under assumptions (B), the binary instrument rule is the sign of the usual Wald estimand. The rule estimated without the use of an IV is the sign of the conditional average treatment effect, a rule relying on the assumption of no unmeasured confounding between the treatment and outcome. For the setting under assumptions (A), the oracle rule is the sign of the conditional average treatment effect for a model controlling for the latent subpopulation types (always-takers, never-takers, and ℓ-compliers), and for the setting under assumptions (B), the oracle rule is the sign of conditional average treatment effect for a model controlling for the unmeasured confounder.

3.3.1 Complier Subpopulation

We generate the instrument $Z \in \{1, 2, 3\}$ using the Categorical distribution with a probability of success of 1/3 for each level. The binary instrument was then generated as 1 if $Z = 3$ and -1 otherwise. We generate two observed variables X_1 and X_2 independently, each following the Bernoulli distribution with probability 1/2 and coded as ± 1. X_1 is chosen to modify the treatment effect and the subpopulation an individual belongs, and X_2 predicts the sub-compliance type. This way, the decision rules can be made to be exactly opposite for the different subpopulations for a given X_1. To satisfy the monotonicity assumption, we first generate whether or not an individual was a complier (either a 2- or 3-complier) with the probability of success $P(A(1) < A(3)|X) = expit(\alpha + 0.25X_1)$, where α is chosen to achieve a predetermined average compliance rate. The average compliance rate denotes the strength of the instrument, where an average closer to 1 implies a stronger IV. Second, if

the individual is not a complier, we generate whether they were a never-taker, as opposed to an always-taker, with probability of success, $P(A(1) = A(2) = A(3) = -1|X) = expit(X_1)$, and if the individual is a complier, we generated whether they were a 2-complier, as opposed to a 3-complier, with probability of success, $P(A(1) < A(2)|X) = expit(2X_2)$. The potential treatments are then defined as $A(1) = A(2) = A(3) = -1$ for never-takers, $A(1) = A(2) = A(3) = 1$ for always-takers, $A(1) = -1$ and $A(2) = A(3) = 1$ for 2-compliers, and $A(1) = A(2) = -1$ and $A(3) = 1$ for 3-compliers. The observed treatment is generated as $A = \sum_{z=1}^{k} \mathbb{1}\{Z = z\}A(z)$. Finally, the outcome is generated as

$$\begin{aligned} Y = 0.5 + 0.5X_1 \\ &+ \Big\{(\delta_{NC} + \Delta_{NC}X_1)(1 + \mathbb{1}\{A(1) = 1\} - \mathbb{1}\{A(3) = 1\}) \\ &+ (\delta_2 + \Delta_2 X_1)(-\mathbb{1}\{A(1) = 1\} + \mathbb{1}\{A(2) = 1\}) \\ &+ (\delta_3 + \Delta_3 X_1)(-\mathbb{1}\{A(2) = 1\} + \mathbb{1}\{A(3) = 1\})\Big\}A + \epsilon, \end{aligned}$$

so that δ_{NC} and Δ_{NC} are the non-compliers' (never-takers and always-takers) treatment parameters, δ_2 and Δ_2 are the 2-compliers' treatment parameters, and δ_3 and Δ_3 are the 3-compliers' treatment parameters. The error term ϵ follows the standard normal distribution. The outcome Y is generated for three different treatment effect scenarios: (1) the decision for each subpopulation is the same, $\delta_{NC} = 0.2$, $\Delta_{NC} = -0.6$, $\delta_2 = 0.3$, $\Delta_2 = -0.8$, $\delta_3 = 0.25$ and $\Delta_3 = -0.7$; (2) the decision for the never-takers and always-takers is exactly opposite from the 2- and 3-compliers, $\delta_{NC} = -0.2$, $\Delta_{NC} = 0.6$, $\delta_2 = 0.3$, $\Delta_2 = -0.8$, $\delta_3 = 0.25$ and $\Delta_3 = -0.7$; and (3) the decision for the 2-compliers is exactly opposite from the never-takers, always-takers and 3-compliers decisions, $\delta_{NC} = 0.2$, $\Delta_{NC} = -0.6$, $\delta_2 = -0.3$, $\Delta_2 = 0.8$, $\delta_3 = 0.25$ and $\Delta_3 = -0.7$. Explicitly, the decision rules for the subpopulations are presented in Table 3.1

Treatment Effect Scenario	Non-complier	2-Complier	3-Complier
(1)	$2(0.2 - 0.6X_1)$	$2(0.3 - 0.8X_1)$	$2(0.25 - 0.7X_1)$
(2)	$2(-0.2 + 0.6X_1)$	$2(0.3 - 0.8X_1)$	$2(0.25 - 0.7X_1)$
(3)	$2(0.2 - 0.6X_1)$	$2(-0.3 + 0.8X_1)$	$2(0.25 - 0.7X_1)$

Table 3.1: The decision rule for each subpopulation in the three Treatment Effect Scenarios.

To identify the optimal ITR, we implement the proposed methods in Section 3.2.3.

Specifically, we use the contrast coefficients $c_z^{(all)}$ to identify the optimal treatment rule for all the ℓ-compliers and the rule $\tilde{d}(X, \hat{L})$ (3.3), that uses a prediction of the latent sub-compliance type and their corresponding ITT effects. To predict the latent sub-compliance type and their ITT effects, we use the contrast coefficients $c_z^{(\ell)}$. The ITT is estimated using a linear model regressing the outcome on the covariates X_1, X_2, the instrument Z, and the interactions between the instrument and covariates. For the binary IV, the ITT was estimated using a similar model but with the binary instrument in place of the multilevel instrument. For the rule $\tilde{d}(X, \hat{L})$, the probability of receiving the treatment is modeled as a logistic regression, regressing the observed treatment on the covariates and their interactions with the instrument. For the rule using no instrument, a linear model regressing the outcome on the covariates and their interactions with the observed treatment is used. For the oracle rule, a linear model regressing the outcome on the covariates, never-takers, 2-compliers, and 3-compliers indicators, and their interactions with the observed treatment is used.

For each scenario, the training sample size is 500 with a test sample size of 10,000. The simulations are repeated 500 times. We evaluate the different decision rules by the empirical meanof the value functions and misclassification rates for the union of all ℓ-compliers, where a larger value function and smaller misclassification rate are preferred. Figure 3.2 presents the results of the simulations.

For Treatment Effect Scenario (1), each of the different decision rules perform well and improve as the average compliance rate increases, where near 0 misclassification rates are achieved and the value functions converge to that of the oracle rule. At lower levels of the compliance rate the rule using no IV performs very well, but this is unsurprising as all of the treatment effects for the different latent subpopulations are very similar and no issues can arise due to weak instruments. The oracle model slightly underperforms for very low compliance rates because some of the subpopulation indicator variables are almost constant, as there are very few compliers, leading to less stable estimation. For Treatment Effect Scenario (2), the decision rules using an instrument continue to perform similarly well. The decision rule ignoring the instrument, however, performs very poorly, particularly at lower compliance rates, where the decision rule conflates the compliers' and non-compliers' effects. Other than the rule $\tilde{d}(X, \hat{L})$ (denoted as $d(X, L)$ in Figure 3.2), the decision rules using an IV all achieve near perfect misclassification rates and converge to the oracle rule's value function. The rule $\tilde{d}(X, \hat{L})$ performs slightly worse than the other rules using an IV in Treatment Effect Scenarios (1) and (2), as it needs to additionally predict the latent sub-compliance type. However, in Treatment Effect Scenario (3), $\tilde{d}(X, \hat{L})$ is able to

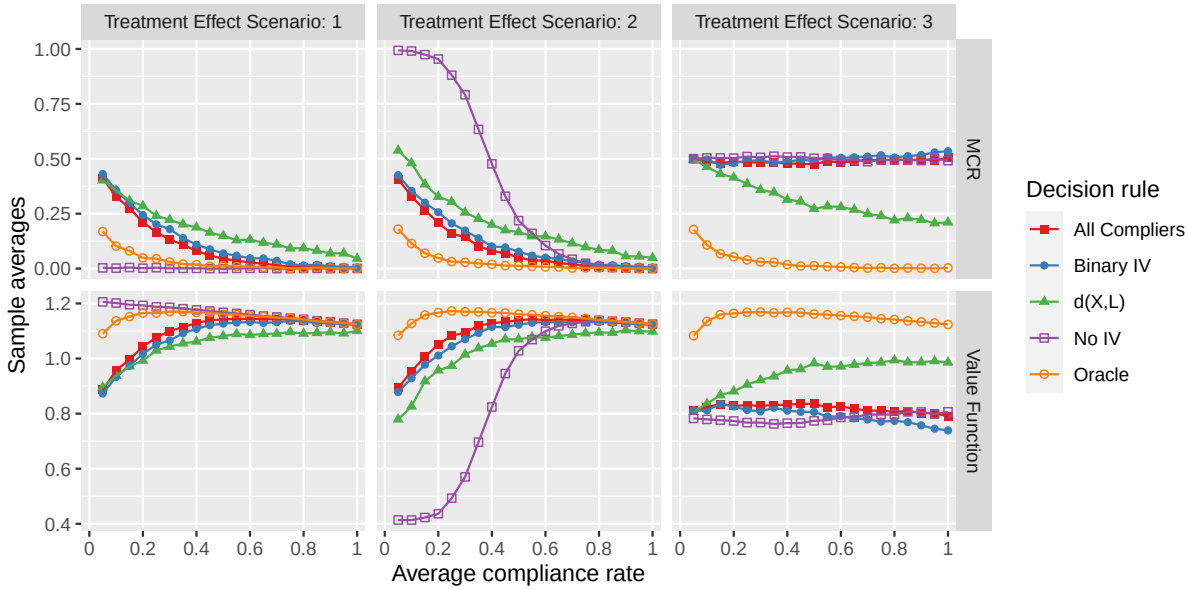


Figure 3.2: The average misclassification rates and value functions of the different decision rules for the complier subpopulation across the three treatment scenarios. The solid squares denote the rule derived using the contrast coefficients $c_z^{(all)}$, the solid circles denote the rule derived for the binary IV, the solid triangles denote the rule $\tilde{d}(X, \hat{L})$, the open squares denote the rule derived using no IV, and the open circles denote the rule derived from the oracle model.

perform well by using their predicted latent sub-compliance types and estimated ITT effects. The other decision rules conflate the two opposite decisions for the 2- and 3-compliers, and are unable to maximize the expected outcome of the union of all ℓ-compliers. We note that the decision rule using the multilevel IV with contrast coefficients $c_z^{(all)}$, denoted as "All Compliers" in Figure 3.2's legend, performs very similarly to the binary IV rule, however, across all settings the rule using the multilevel IV consistently performs better than the binary IV rule across all average compliance rates. This implies that in this setting, dichotimizing the IV still performs very well, but using the IV over its natural support will lead to a more optimal rule.

In Appendix B.9, we present results for additional simulations evaluating the performance for the 2- and 3-complier subpopulations, as well as rules using the contrasts $c_z^{(\ell)}$ separately from $\tilde{d}(X, \hat{L})$. We also present results for a smaller training sample size of 250 and a larger training sample size of 1000. The results are presented in Figures B.2-B.7. The results

in the additional simulations mimic the results presented here, where the methods using IV are able obtain near perfect misclassification rates for the Treatment Effect Scenarios (1) and (2), but failing to maximize the value function for the union of all ℓ-compliers for Treatment Effect Scenarios (3). Only the rule $\tilde{d}(X, \hat{L})$ is able to treat the 2-compliers and 3-compliers separately to increase the value function for the union of all ℓ-compliers in Treatment Effect Scenario (3). In Treament Effect Scenarios (1) and (2), where the 2- and 3-compliers have the same decision rule, the estimated 2- and 3-complier rules, using contrast coefficients, $c_z^{(2)}$ and $c_z^{(3)}$ respectively, perform slightly worse for the entire complier subpopulation, reflecting a loss in efficiency when not combining the two sub-compliance types. The simulations presented in the Appendix B.9 also show a a gradual improvement in misclassification rates and value function as the sample size grows.

3.3.2 Overall Population

We generate the instrument $Z \in \{-1, 0, 1\}$ using the Categorical distribution with a probability of success of 1/4 for levels $Z = -1$ and $Z = 0$ and 1/2 for level $Z = 1$. We change the support of the instrument here for convenience in the data generation process. The binary instrument is then generated as 1 if $Z = 1$ and -1 otherwise. Similar to the simulation settings in Cui and Tchetgen Tchetgen (2021b), we generate five independent Uniformly distributed observed variables over the support $[-1, 1]$ for X and generate the treatment $A \in \{-1, 1\}$ using a logistic regression under three different success probabilities $P(A = 1|X, U, Z)$:

$$
\begin{aligned}
&(1) \quad P(A = 1|X, U, Z) = expit(2X_1 + \beta Z - 0.5U), \\
&(2) \quad P(A = 1|X, U, Z) = expit(2X_1 + \beta Z^2 - 0.5U), \text{ and} \\
&(3) \quad P(A = 1|X, U, Z) = \begin{cases} expit(2X_1 + \beta Z - 0.5U), & X_2 < 0, \\ expit(2X_1 + \beta Z^2 - 0.5U), & X_2 \geq 0, \end{cases}
\end{aligned}
$$

where β is allowed to vary to change the strength of the IV. A larger magnitude of β implies greater IV strength. For $P(A = 1|X, U, Z)$ Scenario (1), the probability of receiving the treatment increases as the instrument increases. For $P(A = 1|X, U, Z)$ Scenario (2), the probability of receiving the treatment increases for values $Z = -1$ and $Z = 1$ of the instrument. For $P(A = 1|X, U, Z)$ Scenario (3), the probability of receiving the treatment is equivalent to Scenario (1) when $X_2 < 0$ and equivalent to Scenario (2) when $X_2 \geq 0$, creating a setting where the IV interacts with X_2 so that some contrasts may

perform better than others. As in Cui and Tchetgen Tchetgen (2021b), the unmeasured U follows a bridge distribution with parameter $\phi = 1/2$ [Wang and Louis (2003)]. The outcome is then generated as $Y = h(X) + q(X)A + 0.5U + \epsilon$, where the main effects and treatment effects are, respectively, $h(X) = 0.5 + 0.5X_1 + 0.8X_2 + 0.3X_3 - 0.5X_4 + 0.7X_5$, and $q(X) = 0.2 - 0.6X_1 - 0.8X_2$, and ϵ follows the standard normal distribution. Under this data generation process, we note that the necessary and sufficient condition is satisfied for both the multilevel and binary instruments.

To identify the optimal ITR, we implement the proposed methods in Section 3.2.4. Specifically, we use the argmax rule, $d_M(X)$ (3.8), which uses the contrasts that maximize the absolute magnitude of the ITT contrast. To derive the argmax contrasts $M(X)$, we consider the following six contrasts: (c1) comparing the lowest level of the IV with the higher two levels, $c_{-1} = -1$, $c_0 = c_1 = 1/2$; (c2) comparing the middle level of the IV with the highest and lowest levels, $c_{-1} = -1/2$, $c_0 = 1$, $c_1 = -1/2$; (c3) comparing the lower two levels of the IV with the highest level, $c_{-1} = c_0 = -1/2$, $c_1 = 1$; (c4) comparing the lowest and middle levels of the IV, $c_{-1} = -1$, $c_0 = 1$, $c_1 = 0$; (c5) comparing the lowest and highest levels of the IV, $c_{-1} = -1$, $c_0 = 0$, $c_1 = 1$; and (c6) comparing the middle and highest levels of the IV, $c_{-1} = 0$, $c_0 = -1$, $c_1 = 1$.

To estimate the ITT, $E[Y|X, Z]$ is estimated by a linear model regressing the outcome on the observed covariates X and their interaction with the multilevel instrument Z. The probability of receiving the treatment, $P(A = 1|X, Z)$, is estimated by a logistic regression, regressing the treatment on the observed covariates and their interaction with the instrument. We then estimate the argmax contrasts $\hat{M}(X)$ by seeing which of the (c1)-(c6) contrasts maximize the absolute value of the ITT contrast for a given X. The argmax rule is estimated as described in Section 3.2.4. Similarly, for the binary IV rule, we use a linear model for the outcome and a logistic regression for the probability of receiving the treatment, each regressing the covariates and their interaction with the binary instrument. To estimate the rule using no IV, we estimate $E[Y|X, A]$ with a linear model regressing the outcome on the covariates and and their interactions with the treatment. For the oracle rule, we estimate $E[Y|X, U, A]$ with a linear model regressing the outcome on the observed and unobserved covariates and the interactions of the observed covariates with the treatment.

For each success probability $P(A = 1|X, U, Z)$ Scenario, the training sample size is 500 with a test sample size of 10,000. The simulations were repeated 500 times. We evaluate the different decision rules by the empirical mean of the value functions and misclassification rates for the overall population. Figure 3.3 presents the results of the simulations.

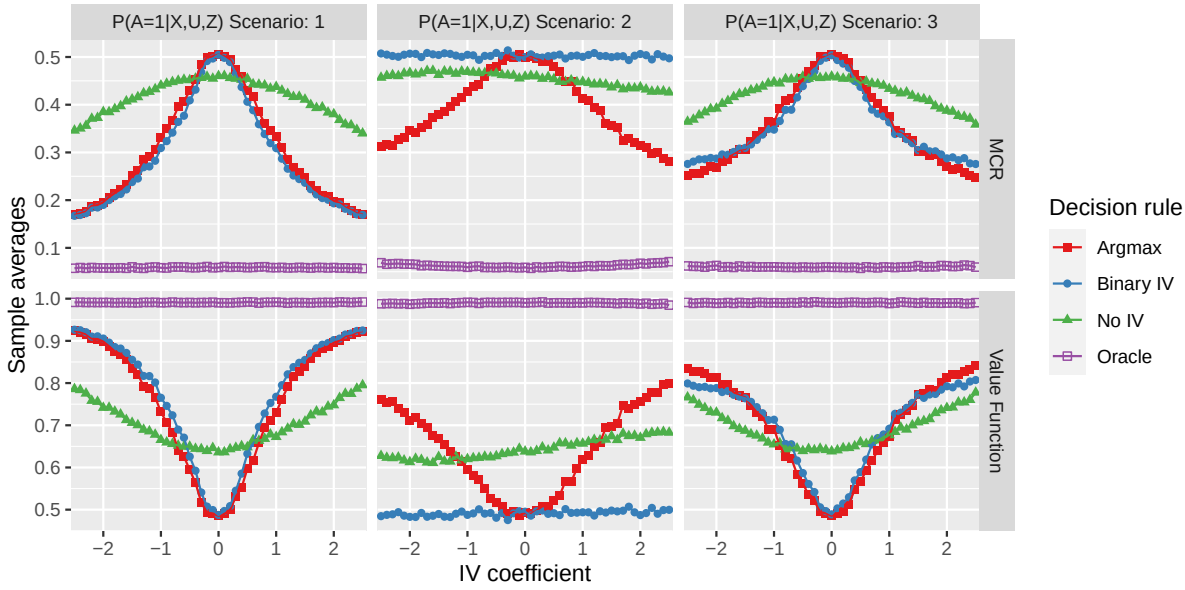


Figure 3.3: The average misclassification rates and value functions of the different decision rules for the overall population across the three $P(A = 1|X, U, Z)$ Scenarios. The solid squares denote the argmax rule using the contrast $M(X)$, the solid circles denote the rule derived for the binary IV, the solid triangles denote the rule using no IV, and the open squares denote the rule derived from the oracle model.

In $P(A = 1|X, U, Z = z)$ Scenario (1), we see both the argmax rule and the rule using a binary IV perform similarly in reducing the misclassification rate and maximizing the value function, particularly at large magnitudes of the IV coefficient β where the instrument is stronger. As the probability of receiving the treatment increases as the instrument increases, dichotomizing the IV does not negate the information provided by the instrument, and so it is not surprising to see the argmax and binary IV rule perform similarly. At smaller IV coefficient magnitudes, we see that the decision rules using the instruments fail to identify an optimal treatment rule, preforming worse than the rule not using an IV. This is due to weak instrument bias. In $P(A = 1|X, U, Z = z)$ Scenario (2), where the lowest level and highest level of the instrument have the same effect on the probability of receiving the treatment, the rule using a binary IV is no longer able to identify an optimal treatment rule. Even at larger magnitudes of the IV coefficient, the rule using the binary IV performs as well as flipping a coin, as seen by a miclassification rate of about 0.5 for all values of the IV coefficient. Though, theoretically, the necessary and sufficient condition is satisfied for a

binary IV in this Scenario, the dichotomization of the multilevel IV by grouping the lower two levels together neutralizes the instrument's effect on the treatment effectively turning a non-weak instrument into a weak instrument. By leaving the multilevel instrument defined over its natural support, the argmax rule is able to use contrasts that do not neutralize the effect of the instrument on the treatment and identify treatment rules that perform much better than the binary IV rule. In $P(A = 1|X, U, Z = z)$ Scenario (3), where the relationship of the instrument on the treatment depends on the value of covariate X_2, the two rules using an IV perform very similarly until larger magnitudes of the IV coefficient when the argmax begins to outperform the rule using the binary IV. The argmax rule begins to achieve greater value than the binary IV rule because it dynamically assigns contrasts to individuals to identify the decision rule depending on the values of X_2. The rule using the binary IV, on the other hand, uses the single difference between levels of the IV, only capturing information that the IV provides when $X_2 < 0$. The rule that does not use an IV performs suboptimally to the rules using the instrument at larger magnitudes of the IV coefficient across all of the success probability $P(A = 1|X, U, Z = z)$ Scenarios, showing that the rule not accounting for the unmeasured confounding on the treatment and outcome fail to identify an optimal treatment rule. At weaker instrument values, the rule using no IV performs as well or better than the rules using IV, however, as the treatment's effect on the outcome is confounded, this result cannot be expected to hold for other settings. Further, as shown in Appendix B.10, when the training sample size grows, the performance of the rules using the IV improves towards the optimal treatment rule while the performance of the rule not using the IV remains the same. Across all of the success probability for the treatment $P(A = 1|X, U, Z = z)$ scenarios, the argmax rule performs well, decreasing the misclassification rate and increasing the overall population's value. The performance of the rules using an IV is largely impacted by the instrument's strength, where performance can vary from the equivalent of flipping a coin to a nearly optimal rule.

In Appendix B.10, we present results for additional simulations evaluating the performance of the argmax rule, a rule using the linear contrast (c5), and a rule using the quadratic contrast (c2) for a smaller training sample size of 250 and a larger training sample size of 1000. The results are presented in Figure B.8 and B.9 of the additional simulations for the overall population. The results in the additional simulations mimic the results presented here, where the rules using an IV outperform the rule not using the IV for all sample sizes across all $P(A = 1|X, U, Z = z)$ Scenarios whenever the IV isn't weak. For the $P(A = 1|X, U, Z = z)$ Scenario (1), the rule using the binary IV and

the argmax rule perform very similarly and improve as the sample size increases. For the $P(A = 1|X, U, Z = z)$ Scenario (2), only the argmax rule's performance approaches the optimal ITR as the sample size grows, as the binary IV rule is still subject to weak instrument bias. In $P(A = 1|X, U, Z = z)$ Scenario (3), the argmax rule has the strongest performance which continues to improve as the sample size grows. The rule using the binary IV generally performs well, but does not see large improvements as the sample size grows. Due to the generation of the multilevel instrument and definition of the binary instrument, the rule using the (c5) linear contrast and the rule using the binary IV perform very similarly in all scenarios. In $P(A = 1|X, U, Z = z)$ Scenario (2), the rule using the (c2) quadratic rule performs best, but suffers in $P(A = 1|X, U, Z = z)$ Scenarios (1) and (3). These results reflect that a single contrast works very well when there is no interaction of the instrument with covariates, as in $P(A = 1|X, U, Z = z)$ Scenarios (1) and (2), but their performance is inferior to the argmax rule in $P(A = 1|X, U, Z = z)$ Scenario (3) when the instrument interacts with covariate X_2.

3.4 Data Analysis

3.4.1 Study Objective and Data Description

We apply our methods to identify optimal ITR for two competing treatments carotid endarterectomy (CEA) and carotid artery stenting (CAS) on treating carotid-artery disease, often referred to as carotid stenosis. The two treatment arms are still regarded as competing despite being well studied, because randomized trials and observational studies disagree in their results as to which procedure is better. Specifically, observational studies suggest that the endarterectomy procedure generally results in better outcomes than the stenting procedure [McPhee et al. (2008); Giles et al. (2010); Wang et al. (2011); Nolan et al. (2012)], whereas randomized trials suggest there is no significant improvement for CEA over CAS [Rosenfield et al. (2016); Brott et al. (2016)]. The current understanding is that the observational studies may be subject to unmeasured selection biases, where sicker, less medically fit patients are disproportionately being selected into the less invasive treatment CAS. As these patients are typically sicker and less medically fit, they are prone to worse outcomes. This selection bias would lead to results showing that CAS is an inferior treatment when variables connected to the selection of treatment are unmeasured. To provide results for an observational study, which may have better external validity, but still control for

unmeasured confounding, studies using an IV are performed, which generally show that there is a smaller difference between the CEA and CAS treatments than presented in the observational studies without an IV [Columbo et al. (2018), Martínez-Camblor et al. (2019a), Martínez-Camblor et al. (2019a)]. The IV used for these studies is center, or hospital, preference and is the same IV used in our analysis.

The data for our analysis was obtained from the Vascular Quality Initiative (`https://www.vqi.org`), a national quality improvement registry containing demographic, clinical, procedural, and outcomes data from over 850 participating centers across the United States and Canada with the purpose of improving vascular care and determining which treatment yields the best results for patients. Specifically, the data included the number of each procedure (CEA and CAS) conducted at the participating centers in the past 6 months prior to the patient's index procedure, an indicator for all-cause mortality or stroke within 30 days of the index procedure, a patient's age, race, sex, and smoking status, whether the patient presented neurologic symptoms, whether the patient elected into treatment, was in need of urgent care, or in need of emergent care, whether the patient was taking any of the preoperative medications, aspirin, P2y12 inhibitors, dual antiplatelets, statins, beta blockers, anticoagulants, or ACE inhibitors, had any of the comorbid conditions, congestive heart failure, coronary revascularization, hypertension, chronic obstructive pulmonary disease (COPD), or COPD treated with home oxygen, or had any prior ipsilateral carotid procedure, prior contralateral carotid procedure, or been diagnosed with coronary-artery disease. Despite such a rich set of covariates, the data are still expected to be exposed to unmeasured selection biases motivating the use of an IV.

Our IV was the center's preference for the treatment CEA and was estimated as the proportion of CEA out of the total number of procedures (CEA and CAS) conducted at each center in the past 6 months prior to a patient's index procedure. The outcome used was the indicator for all-cause mortality or stroke within 30 days of the index procedure. As there are no requirements for a center to join a registry apart of the Vascular Quality Initiative, we excluded hospitals with fewer than 10 total procedures in the 6 months prior to the index procedure. In practice, a recommendation for CEA (or CAS) at a center that performs only CAS (or CEA) would either require the patient to find a different center for their procedure, or the hospital obtain the resources required for the procedure they never perform. These are options that may not be possible if the patient is in need of urgent or emergent care. To focus on the population for which a recommendation can be most practical, we further removed centers that have only conducted one treatment. After the

removal of centers preferring only one treatment and centers conducting fewer than 10 total procedures, we further removed 9,929 patients (about 17%), were lost to follow-up within 30 days of their index procedure and 323 patients (less than 1%) with missing variables and were removed from the analysis. The final data set consisted of 47,930 patients, between 20 and 90 years of age, and 263 centers, where 39,255 patients (about 82%) received the CEA procedure.

3.4.2 Instrument Validity

Before we present the results of our analysis, we discuss the plausibility of the center's preference for treatment CEA as an instrument. We define both a multilevel IV with $k = 3$ levels and a binary IV using the continuous instrument's tertiles and median, respectively. Attending a center with the highest level versus the lowest level of the multilevel instrument resulted in an increased probability of receiving the treatment CEA by about 26%. For the binary IV, the increase in probability of receiving the treatment between the two levels of the IV was about 20%. This satisfies version (B1) of the IV relevance assumption. Though hospital or physician preference is often measured as a surrogate so that the IV is not causally associated with the treatment [Hernán and Robins (2006)], here, the IV is measured as the number of procedures performed in the 6-months prior to the patient's index procedure and so is an estimate of the center's preference. Therefore, the version (A1) of the IV relevance assumption is thought to be satisfied. The (A2/B2) exclusion restriction assumption, states that a center's preference does not affect the outcome except through the procedure performed, and the (A3/B3) IV ignorability assumption, states that the center's preference is unrelated to any unmeasured confounding. Both of these assumptions may be violated, since patients may be aware of what hospital generally performs certain procedures, and they can elect to attend centers that prefer the procedure the patients themselves prefer. Specifically, if some centers have a higher frequency of procedures, it is possible the institution also has a procedure-specific learning effect leading to better outcomes; there has been extensive research on volume-outcome effects in health services literature (see Luft et al. (1979), Halm et al. (2002), and references therein). However, to account for these potential violations of our IV assumptions, we control for the total number of procedures a center has performed in the 6-months prior to the patient's index procedure. This variable is calculated as the sum of total CEA and CAS procedures for the center and including it in our models controls for the possible procedure-specific learning

effect providing plausibility to the exclusion restriction and IV ignorability assumptions.

To identify the value function for the ℓ-complier subpopulations, the (A4) monotonicity assumption is required. This assumption states that no patient who (i) visits a hospital with a higher level of preference for CEA and receives CAS, and (ii) visits a hospital with a lower level of preference for CEA and receives CEA exists in the data. For the multilevel IV, this rules out patients who would receive CEA at level $Z = 1$ but CAS at $Z = 2$ or $Z = 3$ and patients who would receive CEA at level $Z = 2$ but CAS at $Z = 3$. This assumption is plausible, as it follows to reason that if a patient i received CAS ($A_i = -1$) at a center that largely prefers CEA ($Z_i = 3$), then the same patient would have received CAS ($A_i = -1$) at an institution that largely prefers CAS ($Z_i = 1$). However, the monotonicity assumption would be violated if there exist defiers who have variables that lead institutions to select certain treatments against their preference. Swanson and Hernán (2014) give an example of a defier in the similar setting with physician preference as the IV. Consider a patient who is diabetic and physically fit. If this patient were to see a physician who generally prefers the treatment but not for diabetics, or see another physician who generally does not prefer the treatment but makes exceptions for physically fit patients, then this patient would be a defier. Though it is possible that defiers exist, we do not have reason to believe centers have such exceptions to their preferences for the CEA or CAS procedures, and so we assume that monotonicity holds.

To identify the value function for the overall population, the assumption (B4) of no additive U-Z interaction is required. This assumption states that $\delta_c(X) = \tilde{\delta}_c(X, U)$ for all contrasts considered. For this analysis, we consider the same contrasts as in the Section 3.3.2, and that assumption (B4) holds for each. Namely, the contrasts are (c1) comparing the lowest level of the IV with the higher two levels, $c_1 = -1$, $c_2 = c_3 = 1/2$; (c2) comparing the middle level of the IV with the highest and lowest levels, $c_1 = -1/2$, $c_2 = 1$, $c_3 = -1/2$; (c3) comparing the lower two levels of the IV with the highest level, $c_1 = c_2 = -1/2$, $c_3 = 1$; (c4) comparing the lowest and middle levels of the IV, $c_1 = -1$, $c_2 = 1$, $c_3 = 0$; (c5) comparing the lowest and highest levels of the IV, $c_1 = -1$, $c_2 = 0$, $c_3 = 1$; and (c6) comparing the middle and highest levels of the IV, $c_1 = 0$, $c_2 = -1$, $c_3 = 1$. For the binary IV, the assumption of no additive U-Z interaction states that $P(A = 1|X, Z = 1) - P(A = 1|X, Z = -1) = P(A = 1|X, U, Z = 1) - P(A = 1|X, U, Z = -1)$. We cannot know whether an unmeasured variable interacts with center preference, but given we have a rich set of covariates we assume that the assumptions hold. While the assumption of no additive U-Z interaction is needed to hold in order to identify the value function for the overall

population, the necessary and sufficient condition (3.6) needs to be satisfied to identify an optimal ITR for the overall population. That is, an optimal ITR cannot be identified if an unmeasured confounder modifies both the treatment effect and the contrast of the probability of receiving the treatment in a way that the necessary and sufficient condition is violated. As the assumption (B4) of no additive U-Z interaction implies the necessary and sufficient condition holds, we expect to be able to identify an optimal ITR by assuming (B4).

3.4.3 Results

To evaluate the performance of our methods relative to the methods using a binary IV or no IV, we randomly select 75% of patients to train the conditional ITT, probability of receiving the treatment, the interaction of the treatment and outcome ($h_{az}(X)$), and the conditional average treatment effects. The remaining 25% were used as a test set to estimate the value functions. Random forests were chosen for each of these models for its ability to nonparametrically predict the outcomes of interest, its resistance to overfitting, and to allow for many different interaction terms that don't need to be pre-specified in the model. We remark that for each of these models, the R function *randomForest* from version 4.6.14 of the *randomForest* package [Liaw and Wiener (2002)] is used with the default settings. We then apply apply our rule $\tilde{d}(X, \hat{L})$ (3.3) and the rule $d_{(all)}(X)$ (Theorem 3.1) to identify an optimal ITR for the ℓ-compliers, detailed in Section 3.2.3. And, our argmax rule (3.8) over the potential contrasts (c1)-(c6), detailed in Section 3.2.4, is used to identify an optimal ITR for the overall population. For the comparisons using a binary IV, the ITT and the Wald estimand was used to identify an optimal ITR for the ℓ-compliers and the overall population, respectively. For the comparison with the rule using no IV, the conditional average treatment effect was used. To evaluate the performance, we estimated the value function for a given decision rule relative to the value function of the treatment regime of assigning CEA to every patient. That is, we estimated the difference in value functions for a given method and the constant rule of assigning CEA to every patient, $V(d) - V(CEA)$. The value functions were estimated using the empirical forms of (3.5) for the value function of the sub-complier subpopulations, and (3.9) for the value function of the overall population, where the instrument propensity was modeled using a Lasso-penalized multinomial logistic model with shrinkage parameters selected via cross-validation (the default settings of *cv.glmnet* were used for version 4.0.2 of the *glmnet* package [Friedman

et al. (2010)]). The $c_z^{(\ell)}$ contrasts were used to estimate the value for the ℓ-compliers, the $c_z^{(all)}$ contrasts were used to estimate the value for combined group of ℓ-compliers, and the argmax $M(X)$ contrasts were used to estimate the value for the overall population. In Appendix B.11, we present the results using each contrast (c1)-(c6) to estimate the value function of the overall population. We repeated this procedure 200 times. For some of the 200 random splits of the data, in estimating the value for the overall population, the individual value of a few patients was infinite due to the contrast of the probability of receiving the treatment $\delta_c(X) = 0$ in the denominator. In this case, we instead use the truncated mean of the empirical form of (3.9) for more computationally stable results, where 5% of the ends were trimmed. We note that this issue arose in evaluating the performance of the decision rules, so that the estimation of the value for each method was treated similarly.

We present the mean (and standard error) of the difference in value functions across the 200 random splits for the different methods, under the assumptions (A1)-(A4) for the complier subpopulation, in Table 3.2. $V_{(2)}(d)$, $V_{(3)}(d)$, and $V_{(all)}(d)$ denote the estimated value function $V_{(A)}(d)$ using the contrast coefficients $c_z^{(2)}$, $c_z^{(3)}$, and $c_z^{(all)}$ respectively. Given the outcome is an indicator for all-cause mortality or stroke within 30-days, a smaller value is preferred. A negative number have the interpretation of reducing the number of all-cause mortality or stroke within 30-days of the patient's procedure for the different sub-complier subpopulations relative to the treatment strategy of assigning only CEA to patients. A positive value has the opposite interpretation of increasing the number of all-cause mortality or stroke within 30-days relative to the treatment strategy of assigning only CEA to patients.

We see in Table 3.2 that our proposed methods perform best for the 2-complier subpopulation and the combined 2- and 3-complier subpopulation. Particularly, our rules $d_{(all)}(X)$ and $\tilde{d}(X, \hat{L})$ reduces the amount of all-cause mortality and stroke within 30-days by 1.4% and 2% for the 2-compliers, and 0.7% and 0.7% for the combined 2- and 3-compliers, respectively. Our rule appears to be causing harm among 3-compliers however, by increasing the amount of all-cause mortality and stroke within 30-days by 0.8% for the $d_{(all)}(X)$ rule and 1.9% for the $\tilde{d}(X, \hat{L})$ rule. Because both the $d_{(all)}(X)$ rule and the $\tilde{d}(X, \hat{L})$ rule uses the ITT effects of the 2- and 3-compliers, it appears these rules are favoring the 2-compliers over the 3-compliers implying that 2-compliers benefit more from receiving the correct treatment. We remark that a 2-compliers treatment behavior won't change if they attend a center with preference for CEA between the 33% and 66% quantiles, $Z = 2$, or a center with a greater preference for CEA, $Z = 3$, suggesting that their decision rule may be clearer to determine.

The rule using the binary IV and rule using no IV cause less harm for the 3-compliers than our proposed methods, but also provide less benefit for the 2-compliers and combined group of 2- and 3-compliers. This implies that for the more general population of the union of 2- and 3-compliers, using a multilevel IV can result in better decision rules than a binary IV or no IV.

Method	$V_{(2)}(d) - V_{(2)}(CEA)$	$V_{(3)}(d) - V_{(3)}(CEA)$	$V_{(all)}(d) - V_{(all)}(CEA)$
No IV	-0.0045 (0.0132)	0.0041 (0.0295)	-0.0016 (0.0088)
Binary IV	-0.0067 (0.0186)	**0.0013** (0.0454)	-0.0040 (0.0131)
$d_{(all)}(X)$	**-0.0139** (0.0175)	0.0078 (0.0434)	**-0.0068** (0.0121)
$\tilde{d}(X, \hat{L})$	**-0.0196** (0.0176)	0.0185 (0.0422)	**-0.0072** (0.0115)

Table 3.2: Mean (and standard error) of the difference between the estimated sub-complier value functions for the different decision methods and the estimated sub-complier value function for the decision rule of assigning every patient CEA across the 200 random splits of the carotid-artery data. $V_{(2)}(d)$, $V_{(3)}(d)$, and $V_{(all)}(d)$ denote the estimated value function $V_{(A)}(d)$ using the contrast coefficients $c_z^{(2)}$, $c_z^{(3)}$, and $c_z^{(all)}$, respectively.

We present the mean (and standard error) of the difference in value functions across the 200 random splits for the different methods, under the assumptions (B1)-(B4) for the overall population, in Table 3.3. We see that our proposed method performs best for the overall population, as our argmax rule reduces the amount of all-cause mortality and stroke within 30-days by 7.4% for the overall population as compared to the treatment regime of assigning CEA to all patients. For comparison, the rule using the binary IV shows a reduction in the amount of all-cause mortality and stroke within 30-days by 0.4% for the overall population implying that a better rule can be obtained when using a multilevel IV. This is likely due to an improvement in estimation as the binary IV rule may have smaller magnitudes of the differences in the probability of receiving the treatment than the magnitudes of the contrasts of the probability of receiving the treatment used by the argmax rule. The rule not using the IV, which is expected to be confounded by unmeasured variables, appears to cause harm relative to assigning CEA to all patients, where an increase of 0.6% of the amount of all-cause mortality and stroke within 30-days is estimated. In Appendix B.11, where we present the results of the difference in value functions as estimated using single contrasts (c1)-(c6), we see similar results; our proposed argmax rule performs best for

almost all contrasts, though the estimated benefit is more modest when the value functions are estimated using a single contrast.

Method	$V_{(B)}(d) - V_{(B)}(CEA)$
No IV	0.0063 (0.0196)
Binary IV	-0.0039 (0.0296)
Argmax	**-0.0736** (0.0278)

Table 3.3: Mean (and standard error) of the difference between the estimated overall population value function for the different decision methods and the estimated overall population value function for the decision rule of assigning every patient CEA across the 200 random splits of the carotid-artery data.

In a setting where an investigator estimated a single rule to help guide practice, they may want to analyze the covariates to see how the decision rules are recommending CEA versus CAS. By taking the mean of the covariates for the two recommended arms of treatment, an investigator could use an appropriate hypothesis test to see whether the mean of the covariates are dissimilar and gain some insight as to how their decision rule is recommending the two treatments. Such a procedure would provide descriptive statistics of single covariates to see how their decision rule generally prefers assigning one treatment over the other. We perform this procedure for our data set, but because we have 200 random splits of the data, we perform this procedure for each random split and then take the average of the covariates when the hypothesis test deemed the means of the covariates different between the two recommended arms. In particular, we perform this procedure for the age categories to gain insight as to how our rules recommend CEA versus CAS for younger and older patients, as, in general, surgeon's prefer selecting the less invasive treatment CAS for older patients who tend to be less medically fit and the more invasive treatment CEA for younger patients who are tend to be more medically fit. This allows us to determine whether our methods and the rules using the IV follow the surgeon's intuition or not. In Table 3.4, we report the number of random splits (out of 200) the means of the age categories, or the proportions of patients within an age category, were shown to be different by the two-sided Pearson's chi-squared hypothesis test at the level $\alpha = 0.05$. The averages of the proportions of patients within an age category for the different recommended treatment arms is presented in Figure 3.4.

In Figure 3.4, we see that our rules, argmax, $d_{(all)}(X)$, and $\tilde{d}(X, \hat{L})$ agree for all age categories, generally preferring CAS for patients less than 65 and patients greater than

Age Category	Argmax	Wald; Bin.	$d_{(all)}$	$\tilde{d}(X, \hat{L})$	ITT; Bin.	No IV
Less than 65	197	155	193	174	134	200
65 - 69	199	188	198	193	181	85
70 - 74	74	62	78	65	72	78
75 - 79	83	43	76	111	56	164
Greater than 79	187	136	168	190	150	186

Table 3.4: The number of random splits (out of 200) used to calculate the average of the proportions of patients within an age category for the different decision rules. "Argmax" refers to our argmax rule $d_M(X)$, "Wald; Bin." refers to the rule estimated using a binary IV under the set (B) of assumptions, "ITT; Bin." refers to the rule estimated using a binary IV under the set (A) of assumptions, and "No IV" refers to the rule not using an IV.

79, and preferring CEA for patients between 65 and 79 years of age. This can be seen by noting that the proportion of patients is larger for CAS for patients less than 65 and patients greater than 79 for our rules, implying a general preference for CAS for patients within those age categories, and vice versa for patients between 65 and 79 years of age. The rules using the binary IV generally agree with our rules, except to show a smaller difference between the proportion of patients between the ages of 75 and 79. The rule not using an IV, however, generally disagree with the rules using an IV, where the gap between the mean proportions is much larger for patients age 65 or younger and patients between the ages of 75 and 79. Further, the rule not using an IV disagrees with our rules in treatment preference for patients between the ages of 70 and 74, and for patients older than 79 years of age. We believe the rules using an IV better align with the surgeon's intuition and the rule not using an IV better align with the selection bias expected to exist in this data. Other than for patients younger than 65, we see the rules using an IV have a preference for CEA that declines as the patients get older with a change in preference to CAS for patients older than 79 years of age. This reflects the surgeon's preference of assigning the more invasive treatment CEA to younger, more medically fit patients, and the less invasive treatment CAS to older, less medically fit patients. The rule not using an IV, instead, generally prefers to assign CEA to older, less medically fit patients. Such a result reflects a setting in which the treatment effect of CAS is biased by unmeasured variables toward appearing to be a worse treatment than CEA. Older, typically less medically fit

patients are selected to receive CAS but result in poor outcomes, so the rule not using an IV recommends CEA. For patients younger than 65, the results seem to run counter to the surgeon's preference, but we note the gap between the mean proportion of patients is much smaller for the rules using IV than the rule not using IV.

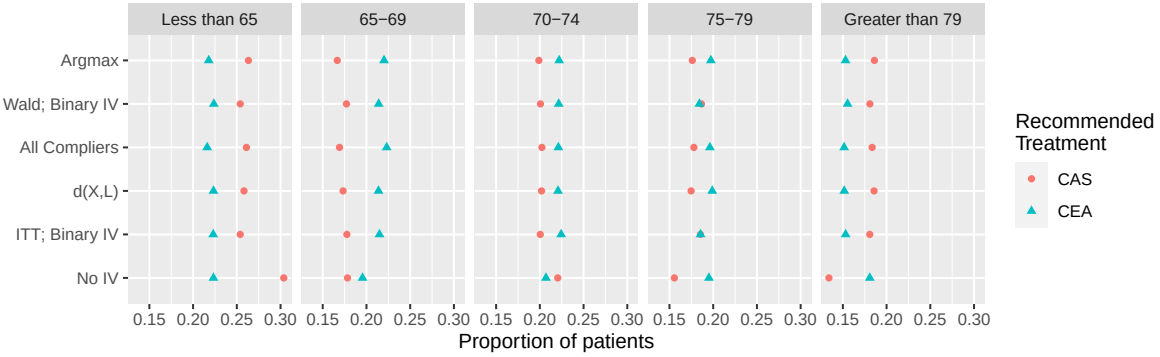


Figure 3.4: The averages of the proportions of patients within an age category for the different recommended treatment arms. The recommended treatment arm that has a larger proportion of patients implies a preference for that treatment for a decision rule. "Argmax" refers to our argmax rule $d_M(X)$, "Wald; Binary IV" refers to the rule estimated using a binary IV under the set (B) of assumptions, "All Compliers" refers to our rule $d_{(all)}(X)$, "d(X,L)" refers to our rule $\tilde{d}(X, \hat{L})$, "ITT; Binary IV" refers to the rule estimated using a binary IV under the set (A) of assumptions, and "No IV" refers to the rule not using an IV.

3.5 Discussion

In this paper, we propose indirect and direct methods to identify optimal ITR in the presence of unmeasured confounding using a multilevel instrument. Our methods can easily be shown to be a generalization of the existing work for IV and ITR for binary IV when the number of levels of the instrument is $k = 2$. We provide similar but distinct approaches to identify optimal ITR for the complier subpopulation under the monotonicity assumption and under the assumption of no additive U-Z interaction in $E[A|X, U, Z]$, and doubly robust AIPW estimates of the value functions under each set of assumptions. Our approaches for identifying optimal ITR contained novel extensions of existing methods. Namely, under monotonicity, we allow for decisions to change based on a latent effect modifier, so long as we can predict this latent effect modifier. And, under the assumption of no additive U-Z

interaction, we use different contrasts of the level of the IV to individualize the estimation of the ITR. We further bridge the two settings of assumptions using necessary and sufficient conditions to describe how the existing unmeasured confounding impacts the identification of an optimal ITR. Our methods were shown to outperform methods using a binary IV in a variety of settings in our simulations. Finally, our methods were also used to identify optimal ITR recommending either CEA or CAS treatment for patients at risk of death or stroke due to carotid-artery disease where we better reduce the number of death or stroke for patients, relative to the rule of assigning CEA to all patients, than the rules using a binary IV.

We conclude by making some recommendations about how to properly use our methods in practice. First, as explained in the introduction, when there is unmeasured confounding present, identifying ITR without IV may provide suboptimal rules. Further, when a non-binary IV is available, which is often the case for health studies, using a binary IV may result in poor estimation of the ITR and value function.

Second, when identifying an optimal ITR for the complier subpopulation under the monotonicity assumption, if there is no specific ℓ-complier subpopulation of particular interest, then we recommend either the rule $d_{(all)}$ using the ITT contrast $c_z^{(all)}$ or the rule $\tilde{d}(X, \hat{L})$ to identify an optimal ITR for the union of all ℓ-compliers. The distinction between using the rule $d_{(all)}$ or $\tilde{d}(X, \hat{L})$ is in the expectation of whether or not ℓ-compliers will have the same or different decisions for a given set of covariate value $X = x$. If an investigator has no prior information on whether the decision rules will agree or disagree for certain covariate values, we recommend using whichever rule $d_{(all)}$ or $\tilde{d}(X, \hat{L})$ results in a better estimated value.

Third, when identifying an optimal ITR for the overall population under the assumption of no additive U-Z interaction in $E[A|X, U, Z]$, we recommend the use of the argmax rule to dynamically use contrasts to estimate the ITR. If an investigator has a preconceived notion of the relationship the instrument has on uptake of the treatment, and that the instrument does not interact much with the observed covariates, then a particular contrast or form of contrasts reflecting this relationship can result in improvement of ITR estimation. That is, we recommend using a single contrast, or limiting the form of contrasts for the argmax rule to consider, if an expert already understands how the instrument affects the probability of receiving the treatment.

Fourth, in light of existing approaches using a binary IV, we observe improvements to using the instrument over its natural support over dichotomizing the IV. However, in

the event that several levels of a multilevel IV, or regions of a continuous IV, have similar effects on the probability of receiving the instrument, efficiency gains in estimation can be attained when grouping these levels together.

4 DISCUSSION AND FUTURE WORK

In this dissertation, we provide several techniques to fill the existing methodology gaps for studying heterogeneous treatment effects with an IV. Specifically, we provide an algorithm to discover novel subgroups and estimate their heterogeneous treatment effects while providing inferential guarantees, and methods to identify optimal ITR using a multilevel IV. We study our methods extensively in simulations and apply our methods to data to study heterogeneous treatment effects of Medicaid and two competing treatments for carotid artery disease.

In Chapter 2, we use IV with matching to detect novel subgroups of compliers and estimate their heterogeneity. Our method first proposes effect modifiers by using an interpretable machine learning technique, namely CART, and second tests whether the proposed effect modifiers are true drivers of heterogeneity through closed testing to strongly control familywise error rate. Further, our method is capable of using the entire data set, avoiding losses in power and efficiency from sample splitting. The simulations demonstrated the method's ability to detect heterogeneous subgroups for varying levels of compliance, where our method was able to generally outperform the BCF-IV method. We applied our method to the OHIE, which used a random lottery as the instrument, to study the heterogeneous effects of Medicaid on the number of days an individual's physical or mental health did not prevent their usual activities. We found that Medicaid benefited complying, older, non-Asian men who selected English materials at lottery sign-up and for complying, younger, less educated individuals who selected English materials at lottery sign-up.

While the work in Chapter 2 aims to provide further methods to study complier treatment effect heterogeneity with an IV, there still remains many avenues of future work. In our algorithm, we allow for any interpretable machine learning technique to be used to suggest potential effect modifiers, but we only study the use of CART. Other techniques less prone to overfitting may lead to a more promising set of effect modifiers. Further, making uninterpretable techniques that are less prone to overfitting more interpretable may also lead to improvements in the true discovery rate and F-score. As our method can incorporate any technique to suggest potential subgroups and still maintain our inferential guarantees, such improvements are a natural direction for improvement. Another possible improvement to our proposed method is the development of different hypothesis tests to use with closed testing. Currently we test whether or not a given subgroup has treatment effects equal to

the hypothesized value, but it may be of more interest to some practitioners to test whether a subgroup's effects are equal to another's, similar to testing whether multiple groups have the same mean in an Analysis of Variance. Such a test is unlikely to suffer in its ability to detect true effect modifiers when treatment effects are opposite but equal.

In Chapter 3, we provide direct and indirect methods using a multilevel IV to identify optimal ITR when the ignorability assumption is assumed to be violated. We use contrasts to generalize from methods using binary IV, which enables the identification of optimal ITR for the latent subpopulations of ℓ-compliers and individualizing ITR estimation for the overall population. We bridge the two settings using slightly different assumptions by providing necessary and sufficient conditions to identify optimal ITR and discussing their similarities. The simulations demonstrated the benefits to estimation by using a multilevel IV over its natural support rather than dichotimizing to a binary IV and suggested how different contrasts can improve ITR estimation in different settings. We applied our methods to the carotid artery data from the Vascular Quality Initiative, using hospital preference as the instrument. We found our methods showed marked improvement over rules using a binary IV and the rule not using an IV, under both settings of assumptions.

While the work in Chapter 3 extends the existing literature using binary IV, more appropriately handling the many common IVs that are not binary in practice, there are several avenues for future work. While we set the framework to use multilevel IV, the first and foremost extension would be to develop methods using a continuous IV to identify ITR. Methods presented in Angrist et al. (2000) and Kennedy et al. (2019) may prove useful to extend our work further to continuous IV. Another extension of our work is to determine optimal ITR using IV for nonbinary treatments. Though no work yet exists studying optimal ITR with IV for nonbinary treatment, there does exist several works extending optimal ITR for a binary treatment to nonbinary IV. Namely, Lou et al. (2018) generalize to a multilevel treatment, Chen et al. (2018) generalize to an ordinal treatment, and Chen et al. (2016) generalize to continuous treatment. As we use contrasts to combine and compare different levels of the multilevel IV, and as contrasts have been used to study heterogeneous treatment effects before [Liang and Yu (2020)], we expect that again contrasts can be used to generalize to multilevel treatment. However, instead of a single vector of contrast coefficients, a matrix of contrast coefficients would be used, where the columns are the contrasts pertaining to the instrument and rows are the contrasts pertaining to the treatment.

Beyond extending to higher dimensions for the instrument and treatment, the work in

Chapter 3 also presents some ideas that are suggestive of possible future work. Namely, we estimate decision rules that map from both observed effect modifiers and latent effect modifiers to the treatment space. Understanding the limits of optimizing the value function when using a latent effect modifier and further methods to use latent effect modifiers are warranted. In Chapter 3, we also provide the argmax rule, which effectively individualizes the estimation of ITR. Conceptually, it is possible some methods work well for certain covariate values, but perform poorly for others. Care would need to be taken to prevent over fitting, but exploring methods to individualize ITR estimation is another avenue of future work.

Finally, in the event that a practitioner believes the ignorability assumption is violated in their data and a valid IV is unavailable, developing sensitivity analyses for estimating heterogeneous treatment effects and identifying optimal ITR is another direction of research. While Kallus and Zhou (2019) and Kallus et al. (2019) present promising techniques, they rely on some assumptions. Namely, they use a marginal sensitivity model that relies on the propensity score to be known or to be consistently estimated. Developing methods under different sensitivity models or sensitivity analyses is warranted.

A APPENDIX FOR CHAPTER 2 "DETECTING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES AND APPLICATION TO THE OREGON HEALTH INSURANCE EXPERIMENT"

A.1 Proof of Proposition 2.1: Familywise Error Rate Control

We recall Proposition 2.1:

Proposition 2.1 (Familywise Error Rate Control of Algorithm 1)**.** *Under the sharp null hypotheses $H_{0\mathcal{L}}$ in Algorithm 1, the conditional probability given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ that the algorithm makes at least one false rejection of the set of hypotheses is at most α.*

Proof. Define $h \subseteq \{1, \ldots, I\}$ to be the union of all groups of pairs for which the null hypothesis is true; the groups of pairs of individuals which have an effect ratio of λ_0. In order to have a notion of Type I error, some hypothesis or hypotheses must be true, so we assume that $h \neq \emptyset$ and that hypothesis $H_{0\mathcal{K}} : \lambda_s = \lambda_0$ for $s = \bigcup_{g \in \mathcal{K}} s_g$ is true. Note that by definition of h, in order for $H_{0\mathcal{K}}$ to be true, the groups in $\mathcal{K}$ are also contained in h, $s \subseteq h$. To make a Type I error and reject hypothesis $H_{0\mathcal{K}}$, Algorithm 1 must reject the intersection of all true hypotheses $H_{0\mathcal{T}}$, where $h = \bigcup_{g \in \mathcal{T}}$ and $\mathcal{K} \subseteq \mathcal{T}$. Yet, rejecting $H_{0\mathcal{T}}$ requires $\left|\frac{T(\lambda_0)}{S(\lambda_0)}\right| \geq z_{1-\alpha/2}$, where $P\left(\left|\frac{T(\lambda_0)}{S(\lambda_0)}\right| \geq z_{1-\alpha/2} \middle| \mathcal{F}, \mathcal{Z}, \mathcal{G}\right) = \frac{\alpha}{2}$. Therefore, to make a Type I error and reject $H_{0\mathcal{K}}$, one must reject $H_{0\mathcal{T}}$ which is a level α test. □

A.2 Additional Simulations: Honest Simultaneous Discovery and Inference

One advantage of our method is being able to use the entire data for discovering and testing effect modifiers. In order to simultaneously discover and draw inference on the sample, we use the absolute value of the pairwise differences $|Y|$ as the outcome of CART to obscure the sign of the difference in adjusted outcomes and preserve the original distribution of the instrumental variables (i.e. distribution based on assumption (A3)). We can then use this distribution to draw inference on our discovered potential effect modifiers. We study this phenomenon in two cases: (1) testing a single effect modifier (i.e. one hypothesis) and (2) testing multiple effect modifiers (i.e. multiple hypothesis).

In the first case, we are concerned about testing a single hypothesis and controlling the Type I error rate after discovering the hypothesis via CART. To investigate the effect of simultaneously discovering and drawing inference, we generate the potential outcomes with no treatment effect, $\lambda_{x_1 x_2} = 0$ for all x_1, x_2, and form potential effect modifiers for two different cases, (i) using $|Y|$ and (ii) Y as the outcome for CART. The first leaf of each tree (or tree's root if no leaves are formed) is then used to test the null hypothesis of no treatment effect. As a result of conducting CART to form a hypothesis 2000 times, Figure A.1 shows the histogram of p-values when we use $|Y|$ as the outcome for CART versus Y as the outcome for CART. Under $|Y|$, the p-values resemble a uniform distribution and hence, Type I error is controlled. However, under Y, the p-values are right-skewed implying that Type I error is inflated. In other words, the null hypothesis is rejected more frequently when using Y as the outcome of CART, demonstrating the "winner's curse" phenomena. Therefore, as predicted by Proposition 2.1, using $|Y|$ as the outcome in CART prevents the contamination of the α level of the hypothesis test and allow for simultaneous discovery and inference.

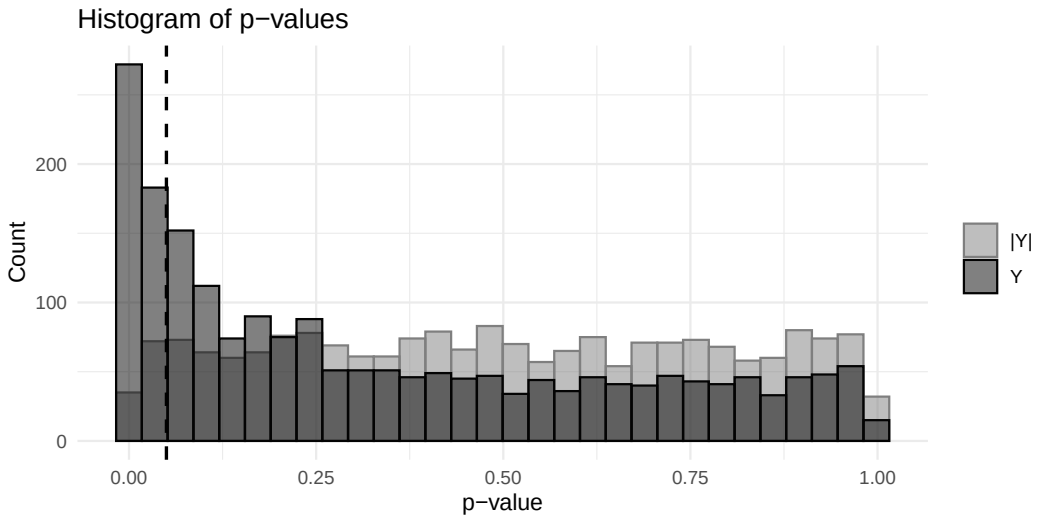


Figure A.1: Histogram of p-values obtained from using $|Y|$ and Y as the outcomes in CART in discovery of potential effect modifiers. The black dashed line denotes the alpha level of 0.05 of the hypothesis tests.

In the second case, with multiple hypotheses, we are concerned with the average Type I error rate and strong control of the familywise error rate rate when testing multiple

Simulation Setting	Mean Type I error rate	Familywise error rate
$\|Y\|$	0.0008	0.028
Y	0.0003	0.028

Table A.1: Results of simulations analyzing strong control of familywise error rate.

hypotheses. To evaluate strong control of the familywise error rate, we generate the data where there is an effect in some groups and no effect in others, $\lambda_{00} = 2$, $\lambda_{01} = \lambda_{10} = 0 = \lambda_{11} = 0$. Furthermore, for *rpart*, we reduce the complexity parameter to 0.0001 to encourage more liberal splitting and set the max depth of the regression tree to be 4 to save on computational time. After data generation we randomly assign instrument values within pairs and split the data using $|Y|$ and Y 2000 times. Each hypothesis is tested for whether or not there is a treatment effect $H_0 : \lambda = 0$, so true hypotheses are hypotheses of groups of pairs generated with $\lambda_{01} = \lambda_{10} = \lambda_{11} = 0$. The average Type I error rate is computed by taking the average of the proportion of false rejections from each of the 2000 simulated trees and the familywise error rate is computed by taking the average of any false rejections among the 2000 simulated trees.

The results of this simulation show that the average type I error rate and familywise error rate are below the α level of the hypothesis tests in both simulation settings $|Y|$ and Y. The average Type I error rate for $|Y|$ and Y is 0.0008 and 0.0003, respectively. The familywise error rate for both $|Y|$ and Y is 0.028 (See Table A.1). This is surprising considering that closed testing requires that each hypothesis test be level α to strongly control the familywise error rate and using Y as the outcome contaminates the test's level. Despite the theoretical underpinnings for this data generation process, closed testing seems to strongly control the familywise error rate regardless of whether or not the test's size is preserved by a technique such as taking the absolute value. Upon closer investigation, it seems that the trees formed in both Y and $|Y|$ cases are the same at the upper levels of the tree, as there is a particularly strong signal for a certain group $\lambda_{00} = 2$. This then leads to the same hypotheses in both settings resulting in similar Type I error rates and familywise error rates. Overall, the simulation suggests that in the case where there is one very strong signal the difference between using $|Y|$ and Y is minor. But, we do stress familywise error control is only guaranteed for the $|Y|$ case.

A.3 Additional Simulations: Detecting H-CACE under Varying Compliance

In Section 2.3, the simulation settings assumed constant compliance rates across the groups. But it is possible that the compliance rates vary between subgroups. Therefore, we further consider four varying compliance rate settings as an extension of understanding the method's performance. These four different compliance settings are referred to as (a) Same, (b) Similar, (c) Different (1), and (d) Different (2) and are functions of the overall compliance rate π. Each are categorized based on the distance from the overall compliance rate. If the overall compliance is less than a half, $\pi \leq 0.5$, a group's compliance rate is $\pi_{x_1x_2} = \pi + c_{x_1x_2}\pi$, and if $\pi > 0.5$, a group's compliance rate is $\pi_{x_1x_2} = \pi + c_{x_1x_2}(1-\pi)$ for some constant $c_{x_1x_2} \in [0,1]$. The four settings are then defined by the constants $c_{x_1x_2}$; (a) Same compliance: $c_{00} = c_{01} = c_{10} = c_{11} = 0$; (b) Similar compliance: $c_{00} = c_{01} = -0.1$ and $c_{10} = c_{11} = +0.1$; (c) Different (1) compliance: $c_{00} = c_{01} = -0.5$ and $c_{10} = c_{11} = +0.5$; and Different (2) compliance: $c_{00} = -0.3$, $c_{01} = -0.5$, $c_{10} = +0.1$, and $c_{11} = +0.7$. When combined with the treatment heterogeneity settings, we have a total of 16 possible settings of heterogeneity in H-CACE. We also remark that subgroups experiencing a larger H-CACE have a small compliance rate. Again, we compare our method to the BCF-IV method (Bargagli-Stoffi et al., 2019).

Figure A.2 shows the true discovery rate of the four compliance types for each treatment heterogeneity setting. As in Section 2.3, this is a measure of the statistical power, where the true discovery rate is defined as the number of false hypotheses rejected out of all false hypotheses tested by closed testing. The different facets denote the treatment effect heterogeneity and the compliance heterogeneity. As the compliance rates get more different, we observe a reduction in the true discovery rate. This is most noticeable in the strong heterogeneity setting, where we see pronounced differences in the true discovery rate among the different compliance settings. In this setting, we also see a change in the true discovery rate between the more different (Different (1) and (2)) and more similar (Same and Similar) compliance groups; the more different compliance groups have an increased true discovery rate after an overall compliance rate of $\pi \approx 0.45$. This is due to the low compliance rate in the groups with stronger H-CACE in Different (1) and Different (2) settings, obscuring the signal for which CART uses to define subgroups. Therefore, the dip described in Section 2.3 and explained in Web Appendix E that is observed in the Strong Heterogeneous and Same Compliance settings (Figure A.2) occurs at a high overall compliance rate for the Different

(1) and Different (2) settings. In the heterogeneity settings, as the overall compliance rate grows, so too does the subgroup-specific compliance rates, and so the true discovery rates of the compliance settings converge. This is evidence that our method's true discovery rate relies on both the subgroup-specific size of H-CACEs and subgroup-specific compliance rates.

Figure A.3 shows the F-score and false positive rate (FPR) of the four compliance types for each treatment heterogeneity setting. As in Section 2.3, these are binary classification measures evaluating the algorithms' ability to predict effect modifiers. For true positives (TP), false positives (FP), and false negatives (FN), the F-score is defined to be $F = \frac{TP}{TP+0.5(FP+FN)}$, and the FPR is defined as the number of false positives out of the negative conditions. Similarly to Figure A.2, as the compliance rates get more different, we observe a reduction in the performance of the algorithms. This is most noticeable in the Slight and Strong Heterogeneity settings for the BCF-IV algorithm and in the Strong and Complex Heterogeneity settings for our proposed algorithm. For the BCF-IV algorithm, the FPR begins to decline at a larger overall compliance rate in the more heterogeneous compliance settings, and the F-score of our algorithm climbs at a later overall compliance rate as well. As was the case with the true discovery rate, this is evidence that the FPR and F-score rely on both the H-CACE and compliance rate heterogeneity.

These simulations are evidence that our method's ability, as measured in statistical power and prediction of effect modifiers relies on both the subgroup-specific size of H-CACEs and the subgroup-specific compliance rates.

A.4 Additional Simulations: Testing for Equal, But Opposite Effects

As it is possible that H-CACEs are equal in magnitude but opposite in direction, it is a question of how our method performs when we take the absolute value of the pairwise differences. To evaluate our method in this specific situation, we generate data as described in Section 2.3 but now considering only the heterogeneity setting $\lambda_{00} = 0.3$, $\lambda_{01} = -0.3, \lambda_{10} = 0.7$ and $\lambda_{11} = -0.7$. Now, there are two effect modifiers x_1 and x_2 where x_1 changes the magnitude of the H-CACE and x_2 changes the direction. We compare our method to the BCF-IV method (Bargagli-Stoffi et al., 2019) as it does not transform the outcome and should still be able to detect the effect modification in this setting.

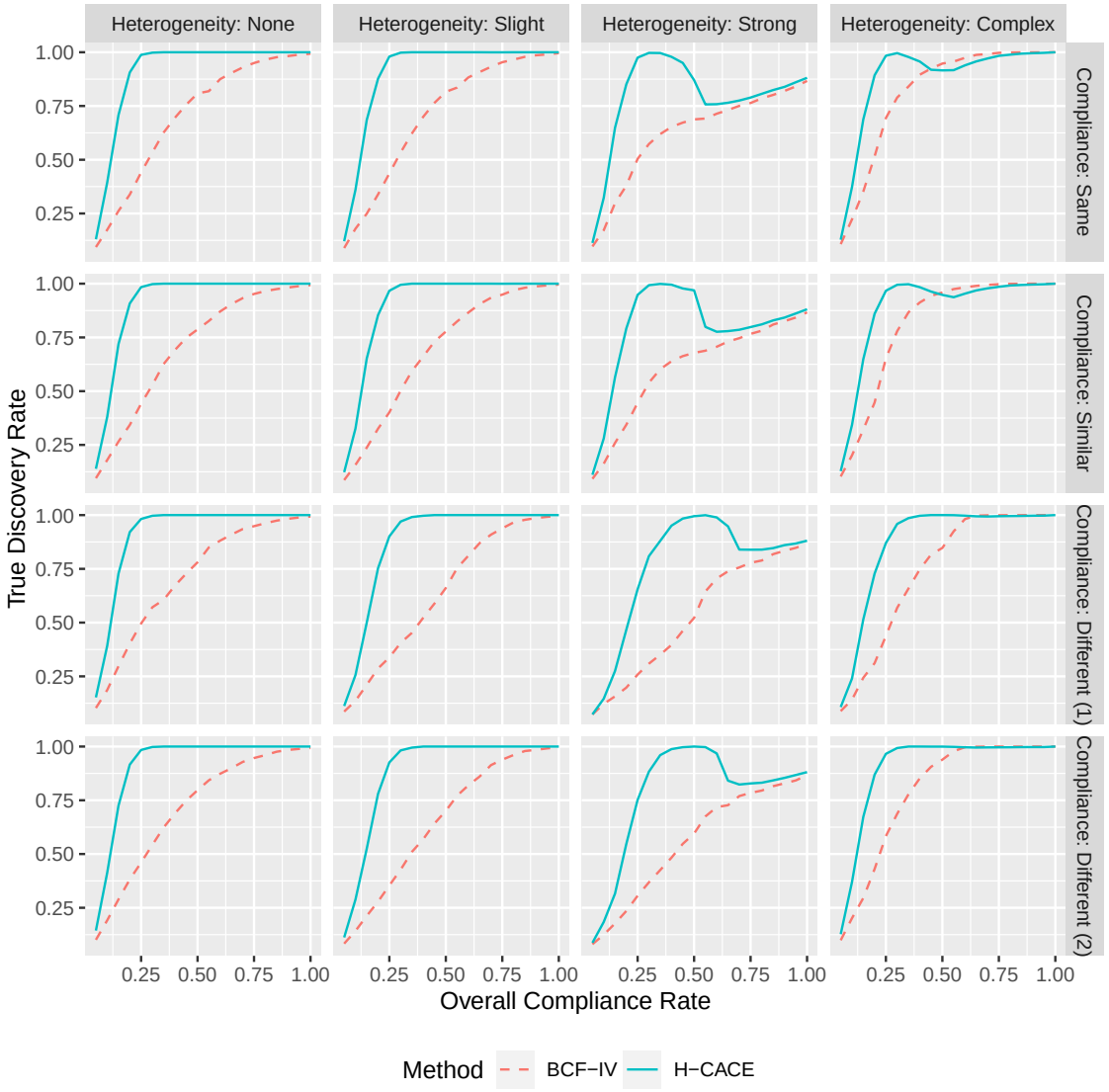


Figure A.2: True discovery rate as a function of overall compliance rate for the four treatment and four compliance heterogeneity settings. The color and line type denote the method, where the red dashed line denotes BCF-IV and the blue solid line denotes our method.

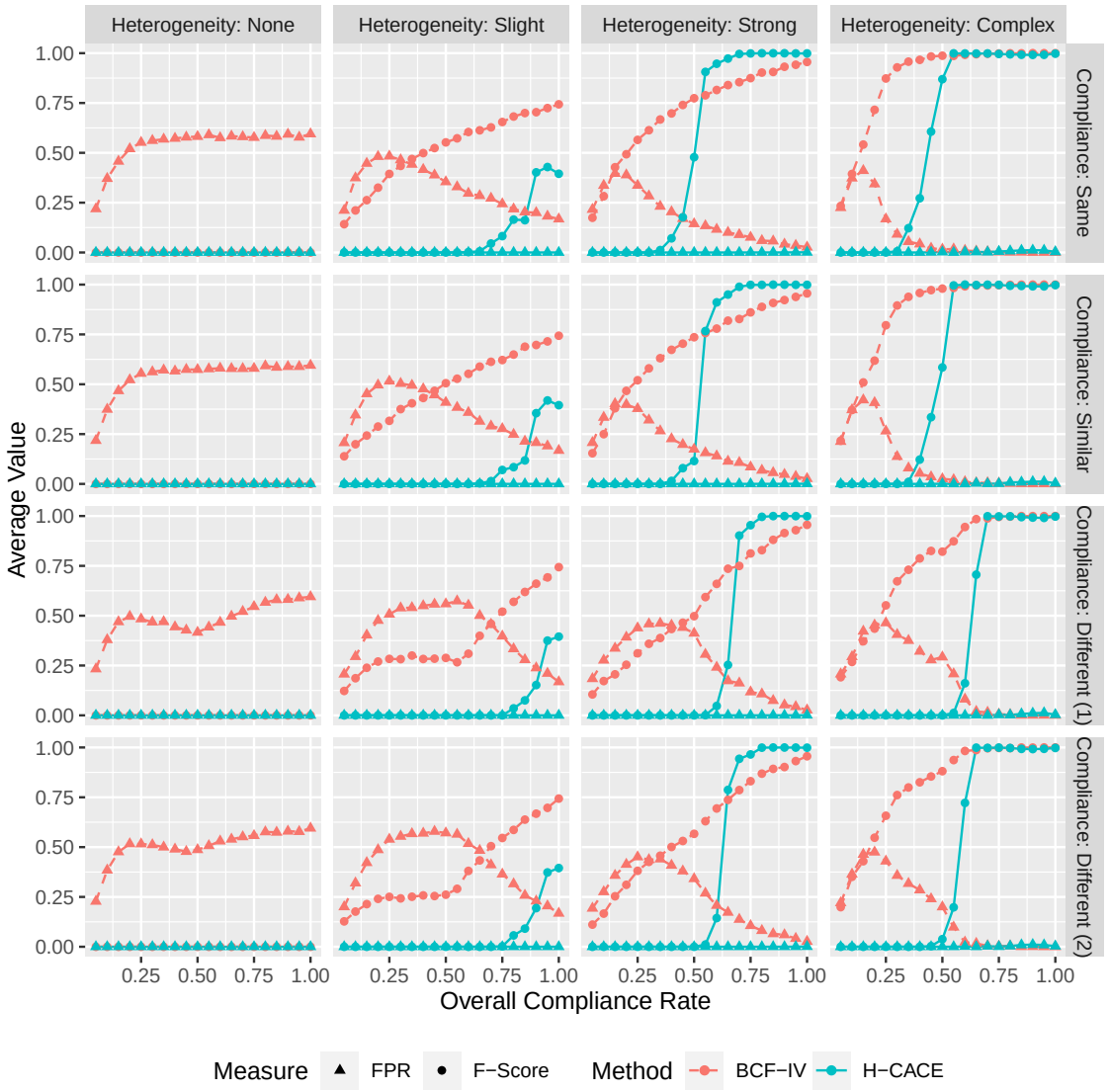


Figure A.3: F-score and false positive rates (FPR) as a function of overall compliance rate for the four treatment and four compliance heterogeneity settings. The shape of the points denote the measure, where a circle denotes the F-score and the triangle denotes the FPR. The color and line type denote the method, where the red dashed line denotes BCF-IV and the blue solid line denotes our method.

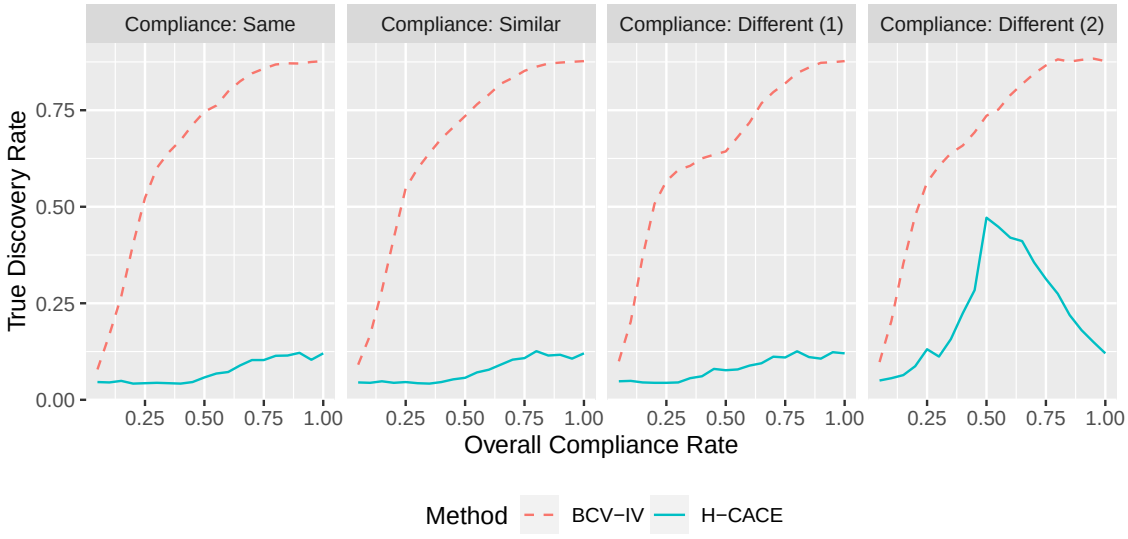


Figure A.4: True discovery rate as a function of overall compliance rate for four compliance heterogeneity settings. The linetype denotes the two methods, where a dashed line denotes the BCF-IV method and the solid line denotes our method.

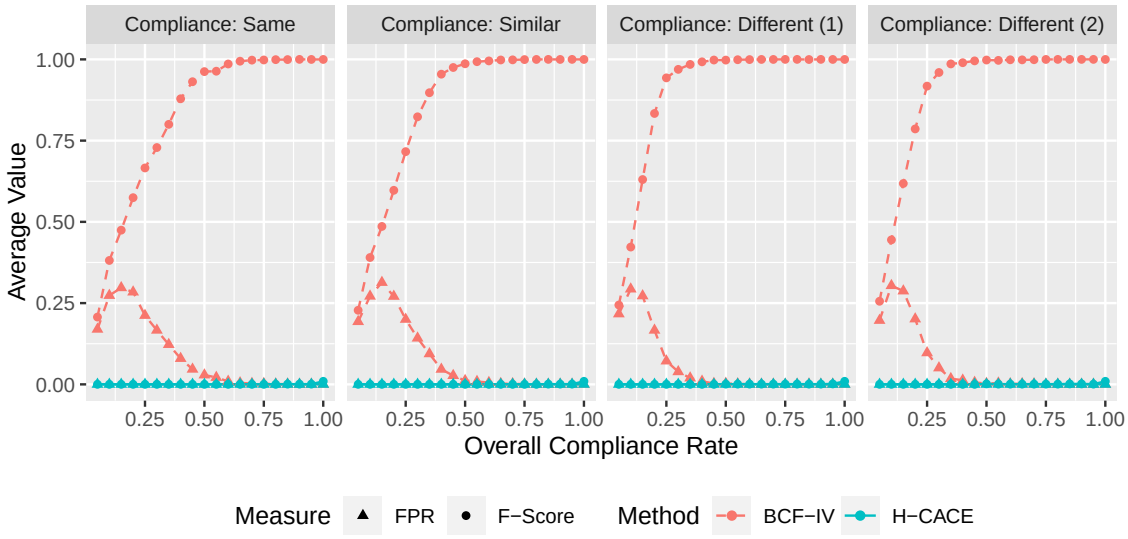


Figure A.5: False positive rate (FPR) and F-score as a function of overall compliance rate for the four compliance heterogeneity settings. The color denotes the two methods, where a red line denotes the BCF-IV method and the blue line denotes our method. The point shapes denote the measure, where a triangle denotes FPR and a circle denotes F-score.

In Figures A.4 and A.5, we see our method underperform in detecting heterogeneous treatment effects in this setting of equal but opposite effects, both in true discovery rate and F-score. The reason is twofold. First, in the transformation of the outcome of CART, the absolute value of the pairwise differences obscures the signal and hinders CART's ability to properly split. Second, in the case that CART does split, our closed testing procedure makes it challenging to reject the hypotheses suggested, since we must first reject the global hypothesis of $H_0 : \lambda_0 = 0$. As we point out in Section 2.3, the CACE λ is a weighted average of the H-CACEs in the global test, and it is therefore unlikely to reject as the average of the H-CACEs is 0. Together, CART's inability to split and the global hypothesis needing to be first rejected, our proposed method underperforms in the setting with equal but opposite effects. We therefore express caution in using our algorithm in the occasion an investigator believes the effect sizes are equal but opposite.

Interestingly, we see a spike of increased true discovery rate for both methods in the Different (2) compliance setting. In this setting, the compliance rates for the four groups are $\pi_{00} = 0.7\pi$, $\pi_{01} = 0.5\pi$, $\pi_{10} = 1.1\pi$, and $\pi_{11} = 1.7\pi$ when the overall compliance rate π is less than or equal to 0.5, and $\pi_{00} = -0.3 + 1.3\pi$, $\pi_{01} = -0.5 + 1.5\pi$, $\pi_{10} = 0.1 + 0.9\pi$, and $\pi_{11} = 0.7 + 0.3\pi$ when $\pi > 0.5$. Therefore, when π is close to 0.5, the compliance rates are approach $\pi_{00} = 0.35$, $\pi_{01} = 0.25$, $\pi_{10} = 0.55$, and $\pi_{11} = 0.85$ for the four groups, changing the weights of the H-CACEs in the average for the global effect and shifting its value away from 0. As the overall compliance approaches 0 or 1, the heterogeneity in the compliance of the four groups reduces so the average of the H-CACEs approaches 0. This all improves our statistical test's ability to reject the global hypothesis and increase the true discovery rate. We note that in the case that the magnitudes are unequal, then our method will return to the performance demonstrated in Section 2.3, as CART will have signal to split on and our global hypothesis will more easily be correctly rejected. This can be seen in an example that the heterogeneity has the form $\lambda_{00} = \lambda_{01} = 0.9$, and the dotted line represents pairs from $\lambda_{10} = \lambda_{11} = -0.1$. With the absolute value transformation of the pairwise differences, our algorithm would treat this setting as in the Strong Heterogeneity setting we simulate in Section 2.3.

A.5 Additional Simulations: Additional Details about the True Discovery Rate

A.5.1 Counter-Intuitive Dip

As mentioned in Section 2.3.1 and shown in Figure A.6, we observe a counter-intuitive dip in true discovery rate as compliance rate grows for our proposed method. To investigate this drop, we also plot the true discovery rate of single subgroups formed by CART. The dashed lines denote leaves containing pairs with a stronger treatment effect and the dotted lines denote leaves containing pairs with a weaker treatment effect. For example, in the strong heterogeneity setting, the dashed line represents pairs from $\lambda_{00} = \lambda_{01} = 0.9$, and the dotted line represents pairs from $\lambda_{10} = \lambda_{11} = 0.1$. For the complex heterogeneity setting, the dashed line denotes pairs generated by $\lambda_{00} = 1.5$; dotted lines aren't shown because CART fails to form a group consisting of only pairs generated by $\lambda_{11} = 0.5$. By comparing the curves, we see that as the compliance rate grows the drop in the true discovery rate is due to the formation of leaves with smaller treatment effects. Because the compliance rate is large enough, these small effects are beginning to be detected by CART. But, the power to detect these effects are much smaller than the large effects, and so the overall true discovery rate, which is roughly the average of these two curves, dips briefly. However, As the compliance rate grows, we see the true discovery rate of our method begin to climb again, as more signal for the smaller H-CACE groups is gained.

A.5.2 Total False Hypotheses

As discussed in Section 2.3.1, the true discovery rate, the number of false null hypotheses rejected divided by the total number of false null hypotheses suggested by the method, is an imperfect measure of the statistical power in some settings of multiple testing with hypotheses generated by a tree-based algorithm. A tree may fail to split resulting in one hypothesis in the denominator of the true discovery rate. If the overall effect is large, this hypothesis test would be rejected and result in a true discovery rate of one. This phenomena is most notable at the strong and complex heterogeneity settings before the dip in the true discovery rate. As shown in Figure A.7, the average number of false hypotheses suggested by our proposed algorithm is close to one at low compliance rates, showing that CART is failing to split. Despite this limitation of the true discover rate as a metric in some settings, our proposed method maintains a high true discovery rate across many settings,

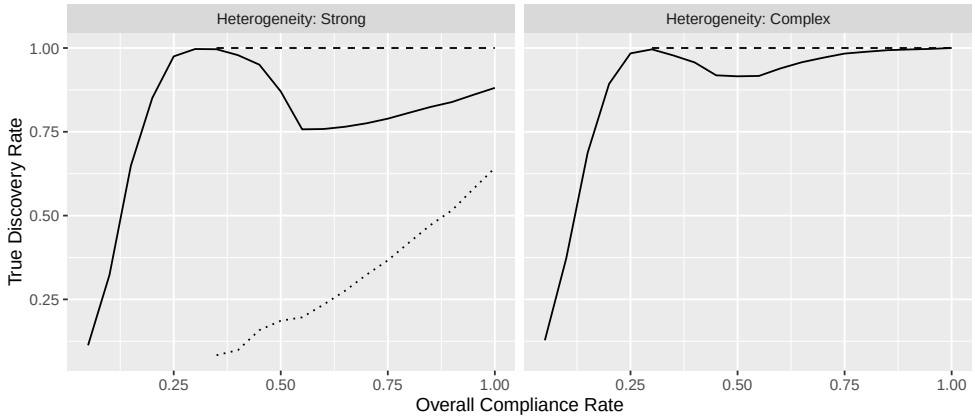


Figure A.6: True discovery rate as a function of overall compliance rate for the strong heterogeneity and complex heterogeneity settings. The line type denotes a single subgroup's treatment effect, where a dashed line denotes the stronger treatment effect and a dotted line the weaker treatment effect.

and suggests that it should be preferred over the competing approach BCF-IV.

A.6 Additional Simulations: Oregon Health Insurance Experiment Semi-Synthetic Simulation

Using the Oregon Health Insurance Experiment (OHIE) as a template for another simulation, we evaluate the performance of the proposed method under treatment magnitudes using the true discovery rate as a measure of the method's statistical power, and false positive rate (FPR) and F-score to measure the method's performance in determining effect modifiers. Using the matched pairs from the OHIE and their pre-instrument covariates $\mathbf{X_{ij}}$, we generate the potential treatments d_{0ij} and d_{1ij} and potential outcomes $r_{0ij}^{(d_{0ij})}$ and $r_{1ij}^{(d_{1ij})}$. Since the design of the OHIE ensures one-sided compliance, we generate the potential treatment without receiving the outcome to also satisfy one-sided compliance, $d_{0ij} = 0$. To have similar compliance rates as observed in Section 2.4, the potential treatment having received the instrument is a Bernoulli trial with success rate $\pi(\mathbf{X_{ij}}) = 0.32 - 0.15(1 - English) + 0.15(English \times Asian) - 0.05(Age < 36)$. Here, *English* is a binary indicator where a value of 1 denotes the individual's preference for English materials in signing up for the lottery, and *Asian* is a binary indicator where a value of 1 denotes the race reported in the survey as Asian. With this heterogeneous compliance rate, the overall compliance is the same as

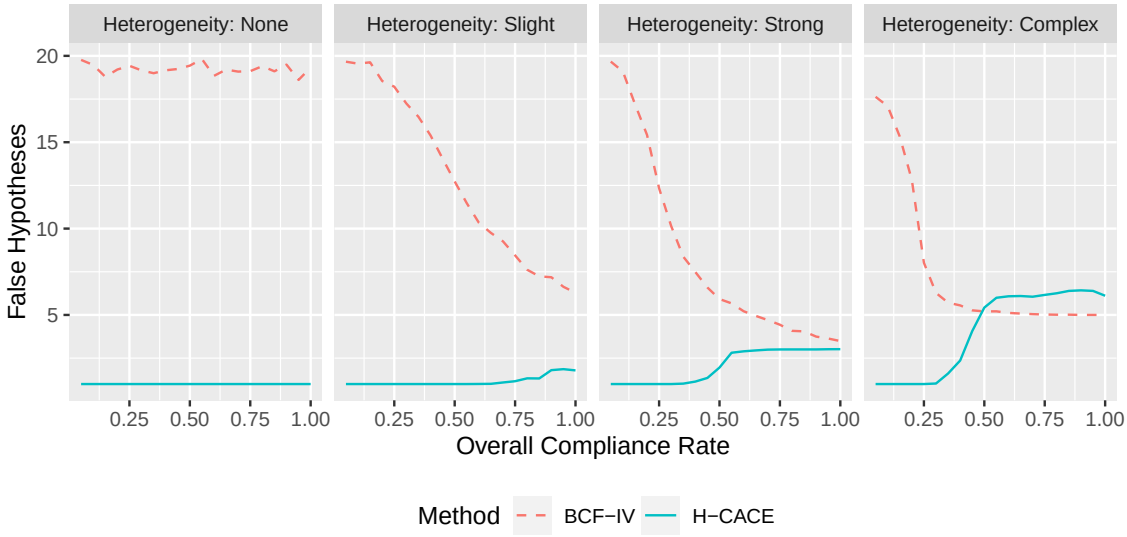


Figure A.7: The average number of false hypotheses suggested by the two algorithms (i.e. the denominator of the true discovery rate) as a function of the compliance rate and heterogeneity settings. The dashed and solid lines denote the BCF-IV procedure and our proposed algorithm, respectively.

that estimated for the sample in Section 2.4, $\pi = 0.29$. We then generate the potential outcomes as in Section 2.3, the potential outcomes having not received the instrument are from a standard normal distribution $r_{0ij}^{(d_{0ij})} \sim N(0,1)$, and the potential outcomes having received the instrument are a function of the H-CACE and effect modifiers detected in Section 2.4, $r_{1ij}^{(d_{1ij})} = r_{0ij}^{(d_{0ij})} + d_{1ij}\lambda(\mathbf{X_{ij}})$. The H-CACE $\lambda(\mathbf{X_{ij}})$ is a function of the effect modifiers *Age*, *Education*, *English*, *Asian*, and *Sex* and the magnitude of the effects are defined in three settings:

(i) Small: $\lambda(\mathbf{X_{ij}}) = 0.25 + 4Age^{-1} + 0.1(1 - Education) - 0.25(1 - English) + 0.35Asian + 0.2(1 - Sex)(Age \geq 36)$

(ii) Moderate: $\lambda(\mathbf{X_{ij}}) = 0.5 + 8Age^{-1} + 0.2(1 - Education) - 0.5(1 - English) + 0.7Asian + 0.4(1 - Sex)(Age \geq 36)$

(iii) Large: $\lambda(\mathbf{X_{ij}}) = 1 + 16Age^{-1} + 0.4(1 - Education) - 1(1 - English) + 1.4Asian + 0.8(1 - Sex)(Age \geq 36)$

As in Section 2.4, *Education* is a binary variable where a value of 1 denotes a vocational degree, 2-year degree, 4-year college degree, or more.

For our proposed method, we use the R package *rpart* with a complexity parameter of 0.001, max depth of 7, and minimum number of observations needed for a split to be 90. This allows CART to split on more variables than the number of effect modifiers while preventing the creation of nodes with very few observations. For BCF-IV, we use the default *rpart* settings as in Section 2.3. The averages of the 1000 simulations at each treatment magnitude level are provided in Table A.2.

The results of this simulation are similar to those seen in Section 2.3, where our proposed method performs well in the true discovery rate, but performs poorly in the FPR and F-score when the compliance rate is low and the heterogeneity magnitudes are weak. This further aligns with our results in Section 2.3, as we found our method struggles to predict effect modifiers as measured by the F-score at compliance rates below 50% and the compliance rate is 29% in this setting. However, we see that our method improves in the F-score as the treatment magnitude increases. In contrast, BCF-IV outperforms our method in the F-score, but has an inflated FPR and a deflated true discovery rate. However, as the signal improves, BCF-IV's FPR reduces to a more ideal value and the true discovery rate grows.

Method	$\lambda(\mathbf{X_{ij}})$	True Discovery Rate	FPR	F-Score
H-CACE	Small	1.00	0.00	0.00
	Moderate	0.99	0.02	0.04
	Large	0.99	0.06	0.76
BCF-IV	Small	0.72	0.17	0.56
	Moderate	0.83	0.05	0.74
	Large	0.93	0.01	0.88

Table A.2: Average true discovery rate, false positive rate (FPR), and F-score of the two methods, H-CACE and BCF-IV, at the different treatment magnitudes.

B APPENDIX FOR CHAPTER 3 "INDIVIDUALIZED TREATMENT RULES WITH MULTILEVEL INSTRUMENTAL VARIABLES"

B.1 Proof of Theorem 3.1

We recall Theorem 3.1:

Theorem 3.1. *Under the assumptions (A1)-(A4), the optimal ITR for ℓ-compliers, or a union of ℓ-compliers, is identified by*

$$\begin{aligned} d^*_{(A)}(X) &= \text{sign}\left\{ E\left[Y(1) - Y(-1) \middle| X, \sum_{z=1}^{k} c_z \mathbb{1}\{A(z) = 1\} = 1 \right] \right\} \\ &= \text{sign}\left\{ \sum_{z=1}^{k} c_z E[Y|X, Z = z] \right\}, \end{aligned}$$

where the contrast $c_z^{(\ell)}$ corresponds to the optimal ITR for the ℓ-compliers, and the contrast for a union of ℓ-compliers can be derived by summing across the ℓ-complier contrasts.

Proof. We have that the ITT contrast is

$$\begin{aligned}
\sum_{z=1}^{k} c_z E[Y|X, Z=z] &= \sum_{z=1}^{k} c_z E\left[\sum_{z=1}^{k} 1\{Z=z\}Y(z, A(z))\middle| X, Z=z\right] \\
&= \sum_{z=1}^{k} c_z E\left[Y\big(z, A(z)\big)|X\right] \\
&= \sum_{z=1}^{k} c_z E\big[1\{A(z)=1\}Y(1) + 1\{A(z)=-1\}Y(-1)\big|X\big] \\
&= \sum_{z=1}^{k} c_z E\big[1\{A(z)=1\}\big(Y(1)-Y(-1)\big)\big|X\big] + \sum_{z=1}^{k} c_z E\big[(Y(-1)\big|X\big] \\
&= E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X\right] \\
&= E\left[Y(1)-Y(-1)\middle| X, \sum_{z=1}^{k} c_z 1\{A(z)=1\}=1\right] \\
&\qquad \times P\left(\sum_{z=1}^{k} c_z 1\{A(z)=1\}=1\middle| X\right) \\
&\quad + E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X, \sum_{z=1}^{k} c_z 1\{A(z)=1\}\neq 1\right] \\
&\qquad \times P\left(\sum_{z=1}^{k} c_z 1\{A(z)=1\}\neq 1\middle| X\right).
\end{aligned}$$

Using the contrast $c_z^{(\ell)}$, defined as $c_{\ell-1}=-1$, $c_\ell=1$, and $c_z=0$ for $z\neq\ell$ for ℓ-compliers, or the sum across the $c_z^{(\ell)}$ contrasts for a union of ℓ-compliers completes the proof. This is seen because $\sum_{z=1}^{k} c_z 1\{A(z)=1\}=0$ for all never-takers, always-takers, and ℓ-compliers not contained in the union of interest, so that

$$\operatorname{sign}\left\{\sum_{z=1}^{k} c_z E[Y|X, Z=z]\right\} = \operatorname{sign}\left\{E\left[Y(1)-Y(-1)\middle| X, \sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right]\right\}.$$

□

B.2 Proof of Proposition 3.1

We recall Proposition 3.1:

Proposition 3.1. *Under the assumptions (A1)-(A4), we have that the decision maximizing* $V(d(X,L))$ *maximizes*

$$\begin{aligned}\arg\max_{d\in\mathcal{D}} V(d(X,L)) &= \arg\max_{d\in\mathcal{D}} E\left[Y\big(d(X,L)\big)\big|A(1)<A(k)\right] \\ &= \arg\max_{d\in\mathcal{D}} E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\big(E[Y|X,Z=\ell]-E[Y|X,Z=\ell-1]\big)\right].\end{aligned}$$

Therefore, if $P(A(\ell)>A(\ell-1)|X)=1$*, for all* ℓ*, we have that*

$$\begin{aligned}d^*(X,L) &= \arg\max_{d\in\mathcal{D}} E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\big(E[Y|X,Z=\ell]-E[Y|X,Z=\ell-1]\big)\right] \\ &= \arg\max_{d\in\mathcal{D}} E\Bigg[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}P(A(\ell)>A(\ell-1)|X) \\ &\qquad\qquad \times\big(E[Y|X,Z=\ell]-E[Y|X,Z=\ell-1]\big)\Bigg] \\ &= \operatorname{sign}\left\{\sum_{\ell=2}^{k}\sum_{z=1}^{k} c_z^{(\ell)}P(A=1|X,Z=z)E[Y|X,Z=z]\right\}.\end{aligned}$$

Proof. We first note that

$$E[Y(d(X,L))|A(k)>A(1)] = \frac{E[Y(d(X,L))(\mathbb{1}\{A(k)=1\}-\mathbb{1}\{A(1)=1\})]}{P(\mathbb{1}\{A(k)=1\}-\mathbb{1}\{A(1)=1\}=1)}$$

As $P(\mathbb{1}\{A(k)=1\}-\mathbb{1}\{A(1)=1\}=1)$ doesn't depend on d, we have that

$$
\begin{aligned}
&E[Y(d(X,L))(\mathbb{1}\{A(k)=1\}-\mathbb{1}\{A(1)=1\})] \\
&\qquad = E\left[Y\big(d(X,L)\big)\left(\sum_{\ell=2}^{k}\mathbb{1}\{A(\ell)=1\}-\mathbb{1}\{A(\ell-1)=1\}\right)\right] \\
&\qquad = E\Bigg[\mathbb{1}\{d(X,L)=1\}\big(Y(1)-Y(-1)\big) \\
&\qquad\qquad \times\left(\sum_{\ell=2}^{k}\mathbb{1}\{A(\ell)=1\}-\mathbb{1}\{A(\ell-1)=1\}\right)\Bigg]+\kappa \\
&\qquad = E\Bigg[\left(\sum_{\ell=2}^{k}\mathbb{1}\{L=\ell\}\mathbb{1}\{d(X,\ell)=1\}\right)\big(Y(1)-Y(-1)\big) \\
&\qquad\qquad \times\left(\sum_{\ell=2}^{k}\mathbb{1}\{A(\ell)=1\}-\mathbb{1}\{A(\ell-1)=1\}\right)\Bigg]+\kappa \\
&\qquad = E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\big(Y(1)-Y(-1)\big)\big(\mathbb{1}\{A(\ell)=1\}-\mathbb{1}\{A(\ell-1)=1\}\big)\right]+\kappa \\
&\qquad = E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\big(\mathbb{1}\{A(\ell)=1\}Y(1)+\mathbb{1}\{A(\ell)=-1\}Y(-1)\big)\right] \\
&\qquad\qquad - E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\big(\mathbb{1}\{A(\ell-1)=1\}Y(1)+\mathbb{1}\{A(\ell-1)=-1\}Y(-1)\big)\right]+\kappa \\
&\qquad = E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\Big(Y\big(\ell,A(\ell)\big)-Y\big(\ell-1,A(\ell-1)\big)\Big)\right]+\kappa \\
&\qquad = E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}E\Big[Y\big(\ell,A(\ell)\big)-Y\big(\ell-1,A(\ell-1)\big)\Big|X\Big]\right]+\kappa \\
&\qquad = E\left[\sum_{\ell=2}^{k}\mathbb{1}\{d(X,\ell)=1\}\Big(E[Y|X,Z=\ell]-E[Y|X,Z=\ell-1]\Big)\right]+\kappa,
\end{aligned}
$$

where $\kappa = E\left[Y(-1)\left(\sum_{\ell=2}^{k}\mathbb{1}\{A(\ell)=1\}-\mathbb{1}\{A(\ell-1)=1\}\right)\right]$ does not depend on d. There-

fore,

$$\arg\max_{d\in\mathcal{D}} V(d(X,L)) = \arg\max_{d\in\mathcal{D}} E\left[\sum_{\ell=2}^{k} \mathbb{1}\{d(X,\ell)=1\}\Big(E[Y|X,Z=\ell] - E[Y|X,Z=\ell-1]\Big)\right].$$

□

B.3 Proof of AIPW Estimator of the Sub-Complier Value Function

Proof. For targeting a specific union of ℓ-compliers, we use the contrast equal to the sum across contrasts $c_z^{(\ell)}$ for the ℓ-compliers contained in the union, so that $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\} = 1$ for all ℓ-compliers contained in the union of interest and $\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\} = 0$ for all ℓ-compliers not contained in the union, never-takers, and always-takers.

Then we have that, under (A4) monotonicity,

$$\begin{aligned}
\sum_{z=1}^{k} c_z P(A=1|Z=z) &= E\left[\sum_{z=1}^{k} c_z \mathbb{1}\{A=1\}|Z=z\right] \\
&= E\left[\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}\right] \\
&= P\left(\sum_{z=1}^{k} c_z \mathbb{1}\{A(z)=1\}=1\right) \\
&:= \delta_c.
\end{aligned}$$

The AIPW estimate is then

$$
\begin{aligned}
&\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{P(Z=z|X)} + h_{az}(X)\right) a\mathbb{1}\{d(X)=a\}\right] \\
&\quad = \sum_{z=1}^{k} c_z E\Bigg\{ \frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right] - h_{az}(X)E\left[\mathbb{1}\{Z=z\}|X\right]\Big) \\
&\qquad\quad + a\mathbb{1}\{d(X)=a\}h_{az}(X)\Bigg\} \\
&\quad = \sum_{z=1}^{k} c_z E\Bigg\{ \frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)}\Big(E\left[\mathbb{1}\{A=a\}Y|X,Z=z\right]P(Z=z|X) - h_{az}(X)P(Z=z|X)\Big) \\
&\qquad\quad + a\mathbb{1}\{d(X)=a\}h_{az}(X)\Bigg\} \\
&\quad = \sum_{z=1}^{k} c_z E\Big\{a\mathbb{1}\{d(X)=a\}\Big(h_{az}(X) - h_{az}(X)\Big) + a\mathbb{1}\{d(X)=a\}h_{az}(X)\Big\} \\
&\quad = \sum_{z=1}^{k} c_z E\left\{a\mathbb{1}\{d(X)=a\}h_{az}(X)\right\} \\
&\quad = E\left\{a\mathbb{1}\{d(X)=a\}\sum_{z=1}^{k} c_z E\Big[\mathbb{1}\{A(z)=a\}Y(z,A(z))\Big|X,Z=z\Big]\right\} \\
&\quad = E\left\{a\mathbb{1}\{d(X)=a\}\sum_{z=1}^{k} c_z E\Big[\mathbb{1}\{A(z)=a\}Y(a)\Big|X\Big]\right\} \\
&\quad = E\left\{a\mathbb{1}\{d(X)=a\}E\left[aY(a)\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}\right)\Bigg|X\right]\right\} \\
&\quad = E\left\{\mathbb{1}\{d(X)=a\}Y(a)\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}\right\} \\
&\quad = E\left\{\mathbb{1}\{d(X)=a\}Y(a)\Bigg|\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right\}P\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right) \\
&\qquad + E\left\{\mathbb{1}\{d(X)=a\}Y(a)\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}\Bigg|\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}\neq 1\right\} \\
&\qquad \times P\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}\neq 1\right).
\end{aligned}
$$

Therefore, using our contrasts for the corresponding target union of ℓ-compliers, we have

$$\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{P(Z=z|X)} + h_{az}(X)\right) a\mathbb{1}\{d(X)=a\}\right]$$
$$= E\left\{Y(d)\middle|\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right\} P\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right).$$

For a misspecified propensity score $\hat{P}(Z=z|X)$, we have

$$\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{\hat{P}(Z=z|X)} + h_{az}(X)\right) a\mathbb{1}\{d(X)=a\}\right]$$
$$= \sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}f(z,X)}{\hat{P}(Z=z|X)}\Big(h_{az}(X) - h_{az}(X)\Big) + a\mathbb{1}\{d(X)=a\}h_{az}(X)\right\}$$
$$= \sum_{z=1}^{k} c_z E\left\{a\mathbb{1}\{d(X)=a\}h_{az}(X)\right\}$$
$$= E\left\{Y(d)\middle|\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right\} P\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right)$$

And for misspecified $\hat{h}_{az}(X)$, we have

$$\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - \hat{h}_{az}(X)\big)}{P(Z=z|X)} + \hat{h}_{az}(X)\right) a\mathbb{1}\{d(X)=a\}\right]$$
$$= \sum_{z=1}^{k} c_z E\Bigg\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right] - \hat{h}_{az}(X)E\left[\mathbb{1}\{Z=z\}|X\right]$$
$$+ P(Z=z|X)\hat{h}_{az}(X)\Big)\Bigg\}$$
$$= \sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right]\Big)\right\}$$
$$= E\left\{Y(d)\middle|\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right\} P\left(\sum_{z=1}^{k} c_z\mathbb{1}\{A(z)=1\}=1\right)$$

□

B.4 Proof of Theorem 3.2

We recall Theorem 3.2:

Theorem 3.2. *Under the set of assumptions (B1)-(B3), the following condition is necessary and sufficient for the sign of the contrast-specific Wald estimand to identify an optimal ITR for the overall population,*

$$E\left[\frac{\tilde{\gamma}(X,U)}{\gamma(X)} \times \frac{\tilde{\delta}_c(X,U)}{\delta_c(X)}\middle| X\right] > 0, \tag{3.6}$$

where $\tilde{\gamma}(X,U) = E[Y(1)-Y(-1)|X,U]$, $\gamma(X) = E[Y(1)-Y(-1)|X]$, $\tilde{\delta}_c(X,U) = \sum_{z=1}^{k} c_z P(A = 1|X,U,Z=z)$, *and* $\delta_c(X) = \sum_{z=1}^{k} c_z P(A=1|X,Z=z)$. *The optimal ITR for the overall population is then*

$$d^*_{(B)}(X) = \text{sign}\left\{E\left[Y(1)-Y(-1)|X\right]\right\} = \text{sign}\left\{\frac{\sum_{z=1}^{k} c_z E[Y|X,Z=z]}{\sum_{z=1}^{k} c_z P(A=1|X,Z=z)}\right\}.$$

Proof. Let $\tilde{\gamma}(X,U) = E[Y(1)-Y(-1)|X,U]$, $\gamma(X) = E[Y(1)-Y(-1)|X]$, $\tilde{\delta}_c(X,U) = \sum_{z=1}^{k} c_z P(A=1|X,U,Z=z)$, and $\delta_c(X) = \sum_{z=1}^{k} c_z P(A=1|X,Z=z)$. The ITT contrast for an arbitrary contrast, such that $\sum_{z=1}^{k} c_z = 0$, under assumptions (B1)-(B3) can be

written as

$$
\begin{aligned}
\sum_{z=1}^{k} c_z E[Y|X, Z=z] &= \sum_{z=1}^{k} c_z E[E(Y|X,U,Z=z)|X] \\
&= E\left[\sum_{z=1}^{k} c_z E\left(\sum_{z=1}^{k} \mathbb{1}\{Z=z\}Y(z,A(z))\middle|X,U,Z=z\right)\middle|X\right] \\
&= E\Bigg[\sum_{z=1}^{k} c_z E\Bigg(\sum_{z=1}^{k} \mathbb{1}\{Z=z\}\Big(\mathbb{1}\{A(z)=1\}Y(1) \\
&\qquad + \mathbb{1}\{A(z)=-1\}Y(-1)\Big)\Bigg|X,U,Z=z\Bigg)\Bigg|X\Bigg] \\
&= E\Bigg[\sum_{z=1}^{k} c_z E\Bigg(\sum_{z=1}^{k} \mathbb{1}\{Z=z\}\Big(\mathbb{1}\{A=1\}Y(1) \\
&\qquad + \mathbb{1}\{A=-1\}Y(-1)\Big)\Bigg|X,U,Z=z\Bigg)\Bigg|X\Bigg] \\
&= E\left[\sum_{z=1}^{k} c_z E\big(\mathbb{1}\{A=1\}\big(Y(1)-Y(-1)\big)\Big|X,U,Z=z\big)\middle|X\right] \\
&\qquad + E\left[\sum_{z=1}^{k} c_z E(Y(-1)|X,U,Z=z)\middle|X\right] \\
&= E\left[\sum_{z=1}^{k} c_z E\big(\mathbb{1}\{A=1\}\big(Y(1)-Y(-1)\big)\big|X,U,Z=z\big)\middle|X\right] \\
&= E\left[E\big\{Y(1)-Y(-1)\big|X,U\big\}\left(\sum_{z=1}^{k} c_z P(A=1|X,U,Z=z)\right)\middle|X\right] \\
&= E\left[\tilde{\gamma}(X,U)\tilde{\delta}_c(X,U)\middle|X\right].
\end{aligned}
$$

Therefore, the contrast-specific Wald estimand can be written as

$$
\frac{\sum_{z=1}^{k} c_z E[Y|X,Z=z]}{\delta_c(X)} = E\left[\frac{\tilde{\gamma}(X,U)\tilde{\delta}_c(X,U)}{\delta_c(X)}\middle|X\right].
$$

Thus, we have that,

$$\frac{\sum_{z=1}^{k} c_z E[Y|X, Z=z]/\delta_c(X)}{\gamma(X)} = E\left[\frac{\tilde{\gamma}(X,U)}{\gamma(X)} \times \frac{\tilde{\delta}_c(X,U)}{\delta_c(X)}\middle| X\right],$$

and so the sign of $\gamma(X)$ necessarily agrees with the conditional contrast-specific Wald estimand whenever the condition above is positive.

□

B.5 Proof of Proposition 3.2

We recall Proposition 3.2:

Proposition 3.2. *Define the contrast $M(X)$ as*

$$M(X) = \arg\max_m \left| \sum_{z=1}^{k} c_z^{(m)} E[Y|X, Z=z] \right|.$$

Under the assumptions (B1)-(B3), we have

$$M(X) = M^*(X).$$

Proof. Define $M(X)$ and $M^*(X)$ as

$$M(X) = \arg\max_m \left| \sum_{z=1}^{k} c_z^{(m)} E[Y|X, Z=z] \right|,$$

and

$$M^*(X) = \arg\min_m \left| E\big[Y(1) - Y(-1)\big|X\big] - \sum_{z=1}^{k} c_z^{(m)} E[Y|X, Z=z] \right|.$$

Note that $M^*(X)$ is equivalent to $\arg\max_m \sum_{z=1}^{k} c_z^{(m)} P(A=1|X,U,Z=z)$. The argument is made by contradiction.

Let $M(X) \neq M^*(X)$, then $\exists$ an $M^\dagger(X) \neq M(X)$ such that

$$\sum_{z=1}^{k} c_z^{(M^\dagger)} P(A=1|X,U,Z=z) > \sum_{z=1}^{k} c_z^{(M)} P(A=1|X,U,Z=z)$$

However, this implies for the ITT contrasts that

$$\left|E\left[E\big[Y(1)-Y(-1)\big|X,U\big]\left(\sum_{z=1}^{k}c_z^{(M^\dagger)}P(A=1|X,U,Z=z)\right)\middle|X\right]\right| > \left|E\left[E\big[Y(1)-Y(-1)\big|X,U\big]\left(\sum_{z=1}^{k}c_z^{(M)}P(A=1|X,U,Z=z)\right)\middle|X\right]\right|$$

and that $M(X) \neq \arg\max_m \left|\sum_{z=1}^{k} c_z^m E[Y|X,Z=z]\right|$. Therefore $M(X) = M^*(X)$. □

B.6 Proof of AIPW Estimator of the Overall Population Value Function

Proof.

$$
\begin{aligned}
&\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y - h_{az}(X)\big)}{P(Z=z|X)} + h_{az}(X)\right)\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\right] \\
&\quad= \sum_{z=1}^{k} c_z E\Bigg\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)\delta_c(X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right] - h_{az}(X)E\left[\mathbb{1}\{Z=z\}|X\right]\Big) \\
&\qquad\quad + \frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\Bigg\} \\
&\quad= \sum_{z=1}^{k} c_z E\Bigg\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)\delta_c(X)}\Big(E\left[\mathbb{1}\{A=a\}Y|X,Z=z\right]P(Z=z|X) - h_{az}(X)P(Z=z|X)\Big) \\
&\qquad\quad + \frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\Bigg\} \\
&\quad= \sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\Big(h_{az}(X) - h_{az}(X)\Big) + \frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\right\} \\
&\quad= \sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\right\} \\
&\quad= E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\sum_{z=1}^{k} c_z E\Big[E\big(\mathbb{1}\{A=a\}Y\big|X,U,Z=z\big)\Big|X\Big]\right\} \\
&\quad= E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\sum_{z=1}^{k} c_z E\Big[E\big(\mathbb{1}\{A=a\}Y(a)\mathbb{1}\{A=a\}\big|X,U,Z=z\big)\Big|X\Big]\right\} \\
&\quad= E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\sum_{z=1}^{k} c_z E\Big[E\big(Y(a)\big|X,U\big)P\big(A=a\big|X,U,Z=z\big)\Big|X\Big]\right\} \\
&\quad= E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}E\Big[E\big(Y(a)\big|X,U\big)a\tilde{\delta}_c(X,U)\Big|X\Big]\right\} \\
&\quad= E\left\{\frac{\mathbb{1}\{d(X)=a\}}{\delta_c(X)}E\Big[E\big(Y(a)\big|X,U\big)\delta_c(X)\Big|X\Big]\right\} \\
&\quad= E\left\{E\Big[\mathbb{1}\{d(X)=a\}Y(a)\Big|X\Big]\right\} \\
&\quad= E\left\{Y(d)\right\}
\end{aligned}
$$

For a misspecified propensity score $\hat{P}(Z=z|X)$, we have

$$\begin{aligned}
&\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y-h_{az}(X)\big)}{\hat{P}(Z=z|X)}+h_{az}(X)\right)\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\right] \\
&\quad=\sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}f(z,X)}{\hat{P}(Z=z|X)\delta_c(X)}\Big(h_{az}(X)-h_{az}(X)\Big)+\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\right\} \\
&\quad=\sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}h_{az}(X)\right\} \\
&\quad=E\left\{Y(d)\right\}.
\end{aligned}$$

And for misspecified $\hat{h}_{az}(X)$, we have

$$\begin{aligned}
&\sum_{z=1}^{k} c_z E\left[\left(\frac{\mathbb{1}\{Z=z\}\big(\mathbb{1}\{A=a\}Y-\hat{h}_{az}(X)\big)}{P(Z=z|X)}+\hat{h}_{az}(X)\right)\frac{a\mathbb{1}\{d(X)=a\}}{\delta_c(X)}\right] \\
&\quad=\sum_{z=1}^{k} c_z E\Bigg\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)\delta_c(X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right]-\hat{h}_{az}(X)E\left[\mathbb{1}\{Z=z\}|X\right] \\
&\qquad\quad+P(Z=z|X)\hat{h}_{az}(X)\Big)\Bigg\} \\
&\quad=\sum_{z=1}^{k} c_z E\left\{\frac{a\mathbb{1}\{d(X)=a\}}{P(Z=z|X)\delta_c(X)}\Big(E\left[\mathbb{1}\{Z=z\}\mathbb{1}\{A=a\}Y|X\right]\Big)\right\} \\
&\quad=E\left\{Y(d)\right\},
\end{aligned}$$

where the last equality follows from a similar argument as shown in Cui and Tchetgen Tchetgen (2021b) and the fact that $\sum_{z=1}^{k} c_z = 0$. □

B.7 Proof of Necessary and Sufficient Conditions Under Assumption Set (A)

Proof. Under the assumptions (A1)-(A4), we have from the proof of Theorem 3.1 that

$$\sum_{z=1}^{k} c_z E[Y|X,Z=z] = E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X\right]$$

Therefore, defining L as the latent sub-compliance type, we have

$$\begin{aligned}
\sum_{z=1}^{k} c_z E[Y|X, Z=z] &= E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X\right] \\
&= E\left\{E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X, A(1)<A(k)\right]\middle| X\right\} \\
&= E\Bigg\{\sum_{\ell=2}^{k} E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X, L=\ell\right] \\
&\qquad \times P\big(L=\ell|X, A(1)<A(k)\big)\Bigg| X\Bigg\}
\end{aligned}$$

We only now use the assumption (A4) monotonicity, so that when using the contrast coefficients $c_z^{(\ell)}$ so that $\sum_{z=1}^{k} c_z 1\{A(z)=1\}=1$ only for the ℓ-compliers and is equal to 0 for all other subpopulations. We then have that

$$\begin{aligned}
\sum_{z=1}^{k} c_z E[Y|X, Z=z] &= E\Bigg\{\sum_{\ell=2}^{k} E\left[\big(Y(1)-Y(-1)\big)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X, L=\ell\right] \\
&\qquad \times P\big(L=\ell|X, A(1)<A(k)\big)\Bigg| X\Bigg\} \\
&= E\left\{E\left[Y(1)-Y(-1)|X, L\right] P\left(\sum_{z=1}^{k} c_z 1\{A(z)=1\}=1\middle| X, A(1)<A(k\right)\middle| X\right\}
\end{aligned}$$

Thus, we have that

$$\frac{\sum_{z=1}^{k} c_z E[Y|X, Z=z]/\delta_c(X)}{\gamma(X)},$$

is equivalent to

$$E\left\{\frac{E\left[Y(1)-Y(-1)|X, L\right]}{\gamma(X)} \times \frac{P\left(\sum_{z=1}^{k} c_z 1\{A(z)=1\}=1\middle| X, A(1)<A(k)\right)}{\delta_c(X)}\middle| X\right\}$$

and so the sign of $\gamma(X)$ necessarily agrees with the conditional Wald estimand using contrast coefficients $c_z^{(\ell)}$ whenever the above expression above. Without the monotonicity assumption,

the necessary and sufficient condition has the form

$$E\left\{\frac{\sum_{\ell=2}^{k} E\left[\left(Y(1)-Y(-1)\right)\sum_{z=1}^{k} c_z 1\{A(z)=1\}\middle| X, L=\ell\right] P(L=\ell|X, A(1)<A(k)}{\gamma(X)\delta_c(X)}\middle| X\right\} > 0$$

for the set (A) of assumptions.

□

B.8 Additional Simulations: Estimating the Value Function with the Argmax Contrasts

We conduct additional simulations to evaluate the consistency of using the argmax contrasts to estimate the value function of the overall population. The settings for this simulation study are the same as those described in Section 3.3.2, except we use an observed IV so that the instrument propensity depends on the observed covariates X. Otherwise, we generate the observed covariates, the unobserved covariates U, treatment A, and the potential outcomes $Y(-1)$ and $Y(1)$ as described in Section 3.3.2. For these simulations, we consider the success probability $P(A=1|X,U,Z)$ Scenario (1). The observed IV, $Z \in \{-1,0,1\}$, follows a multinomial distribution with instrument propensities $P(Z=-1|X) = 1/(1+e^{X\beta_0}+e^{X\beta_1})$, $P(Z=0|X) = e^{X\beta_0}/(1+e^{X\beta_0}+e^{X\beta_1})$, and $P(Z=1|X) = e^{X\beta_1}/(1+e^{X\beta_0}+e^{X\beta_1})$, where the regression coefficients are different for the different levels of the IV; for $Z=0$ the regression coefficient is $\beta_0 = (1/2, 1/2, 0, 0, 0)'$, and for $Z=1$ the regression coefficient is $\beta_1 = (1,1,0,0,0)'$. To derive the decision rules we use our argmax rule, the linear contrast $c_{-1}=-1$, $c_0=0$, and $c_1=1$, and the quadratic contrast $c_{-1}=-1/2$, $c_0=1$, and $c_1=-1/2$, where the models used are the same as described in Section 3.3.2. To evaluate the consistency of using the argmax contrasts to estimate the value function, we estimate the value function using the the AIPW estimator (3.9) with the argmax contrasts $M(X)$ and compare it with the empirical mean of the value function. We use a multinomial model regressing the IV on the observed covariates to estimate the IV propensity scores and linear models to estimate h_{az} for the AIPW estimates. We perform these simulations 500 times at varying training sample sizes for the models and present the results in Figure B.1.

The results shown in Figure B.1 imply that using the argmax contrasts with the AIPW estimator can consistently estimate the value function for larger magnitudes of the IV

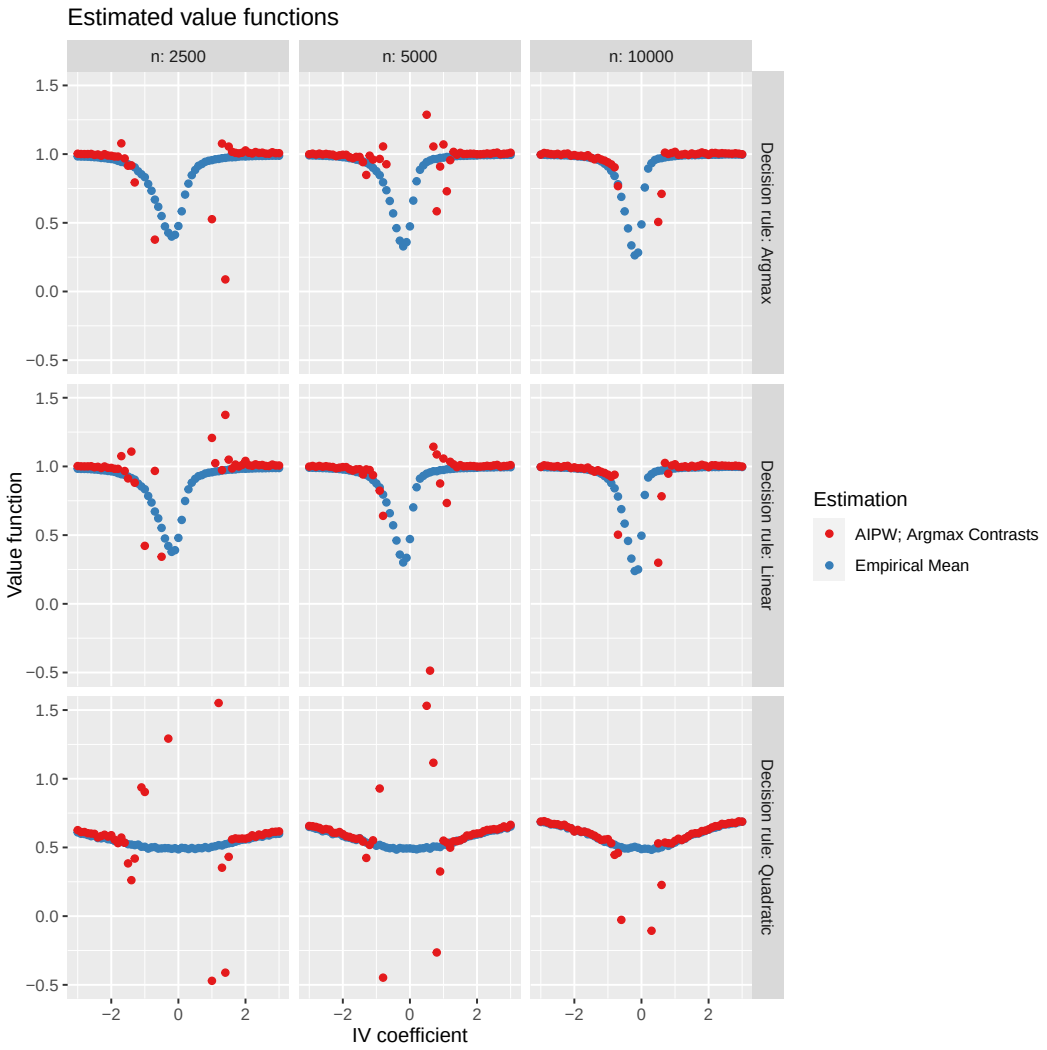


Figure B.1: The average estimates of the overall population value function for the argmax rule, a rule using a linear contrast, and a rule using a quadratic contrast. The red points denote the AIPW estimates using the argmax contrasts and the blue points denote the empirical mean.

coefficient. This can be seen by the mean of the AIPW estimates with the argmax contrast overlapping the empirical mean of the value function for all three decision rules at larger magnitudes of IV coefficients. For IV coefficients close to 0, the instrument becomes too weak for the value function to be consistently estimated, however, the weak instrument bias can be mitigated as the training sample size increases.

B.9 Additional Simulations: Sub-Complier Subpopulations

We conduct additional simulations under the set (A) of assumptions. The settings from the original simulations in Section 3.3.1 are unchanged, except to include sample sizes of 250 and 1000, and consider rule $d_{(2)}(X)$ using the $c_z^{(2)}$ contrast to identify the optimal ITR for 2-compliers, and the rule $d_{(3)}(X)$ using the $c_z^{(3)}$ contrast to identify the optimal ITR for 3-compliers,. Figure B.2 presents the misclassification rates of the different decision rules for the union of both 2- and 3-compliers and Figure B.3 presents the value functions of the different decision rules for the union of both 2- and 3-compliers.

The results in Figures B.2 and B.3 mimic those seen in Section 3.3.1, where all rules using an IV generally perform well for Treatment Scenarios (1) and (2), but only the rule $\tilde{d}(X, \hat{L})$ performs well in Treatment Scenario (3) when the 2- and 3-compliers have opposite decision rules for the different values of X_1. What we see from these results is the gain in performance as the sample size increases for all rules, as well as how rules targeting the 2- and 3-compliers perform. Given that the rules for the 2- and 3-compliers agree for Treatment Effect Scenarios (2) and (3), it is unsurprising to see both of the rules $d_{(2)}(X)$ and $d_{(3)}(X)$ reducing the misclassification rate and maximizing the value function as the average compliance rate grows. Given that there are fewer 2- and 3-compliers separately, it is also unsurprising to see the rule $d_{(all)}(X)$ using the contrast $c_z^{(all)}$ outperform $d_{(2)}(X)$ and $d_{(3)}(X)$. As the rule using $d_{(all)}(X)$ pools together the 2- and 3-compliers, it gains efficiency when the decision rules of 2- and 3-compliers agree.

Figure B.4 presents the misclassification rates of the different decision rules for the 2-compliers and Figure B.5 presents the value functions of the different decision rules for the 2-compliers. The misclassification rates and value function results for the 2-compliers mirror those seen for the union of all ℓ-compliers for Treatment Effect Scenarios (1) and (2), which is expected since the decision rules for the 2- and 3-compliers are the same in those Scenarios. For Treatment Effect Scenario (3), however, we see that $d_{(2)}(X)$ performs best as it focuses only on estimating the treatment decision for the 2-compliers. Further, $d_{(3)}(X)$

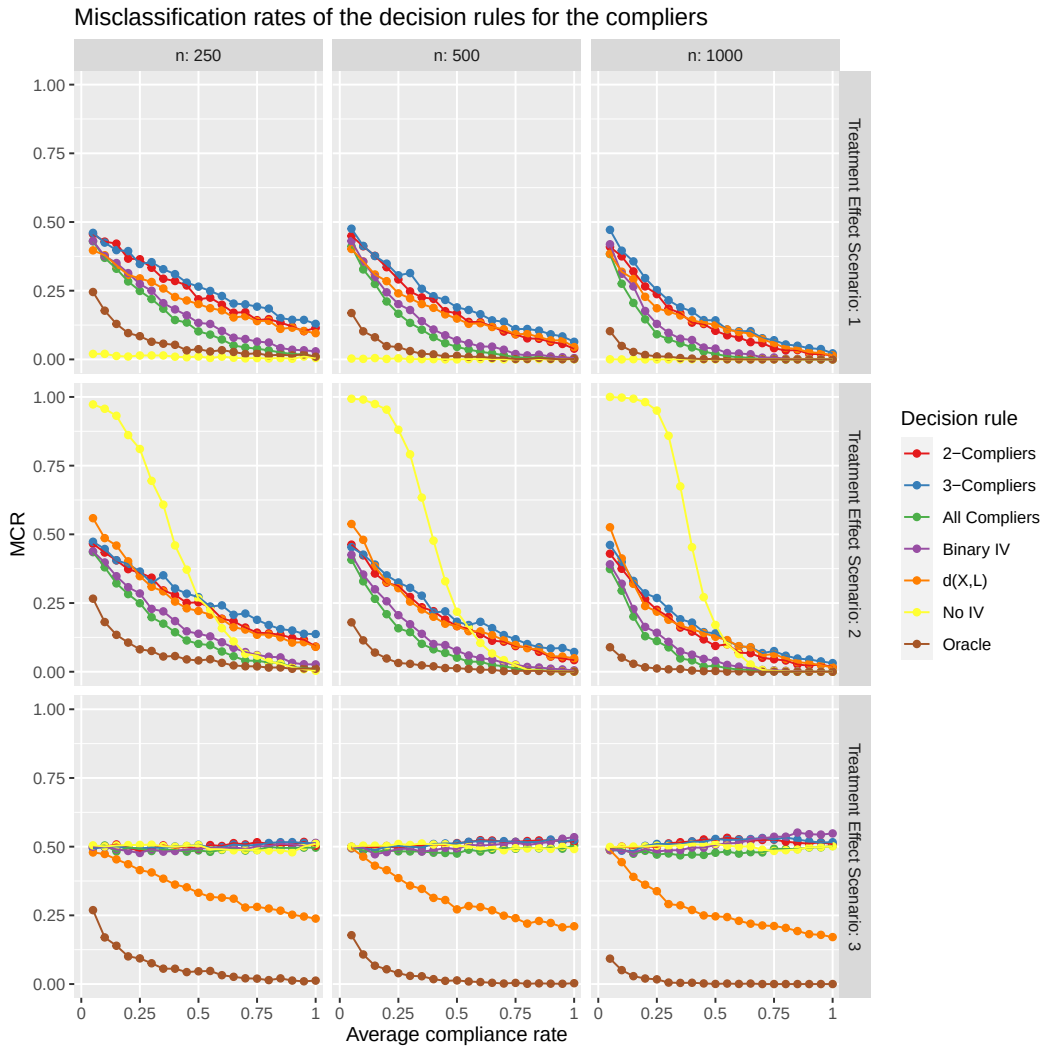


Figure B.2: The average misclassification rates of the different decision rules for all ℓ-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

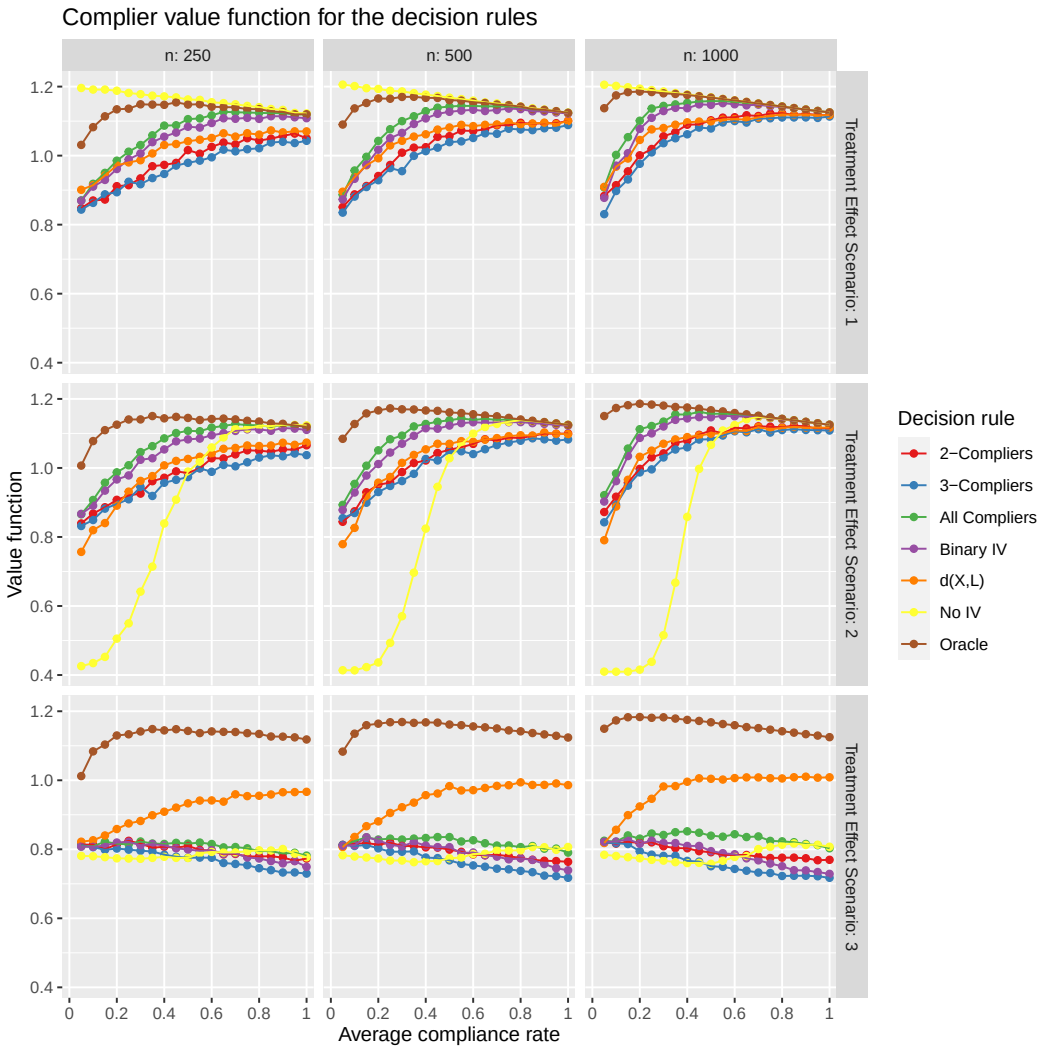


Figure B.3: The average value functions of the different decision rules for all ℓ-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

and the rule using the binary IV cause harm as seen by the increasing misclassification rate and decreasing value function as the average compliance rate grows. This is unsurprising for $d_{(3)}(X)$ considering it focuses on estimating the ITR for the 3-compliers, who have the opposite rule to the 2-compliers. The binary IV appears to be favoring the 3-compliers over the 2-compliers in estimation of the ITR resulting in poor performance for the 2-complier subpopulation.

Figure B.6 presents the misclassification rates of the different decision rules for the 3-compliers and Figure B.7 presents the value functions of the different decision rules for the 3-compliers. The misclassification rates and value function results for the 3-compliers continue to mirror those seen for the union of all ℓ-compliers for Treatment Effect Scenarios (1) and (2), as the decision rules for the 2- and 3-compliers are the same. For Treatment Effect Scenario (3), we see the reverse of what is shown in Figures B.4 and B.5 for the 2-compliers. Namely, $d_{(3)}(X)$ now performs well whereas $d_{(2)}(X)$ and the binary IV rule perform poorly. The rule using no IV performs very well in Treatment Effect Scenario (3) at lower average compliance rates. This is because the non-compliers and 3-compliers have the same decision rules for this Scenario, and at low average compliance rates there are few compliers and therefore few 2-compliers. As the number of 2-compliers in the data increases, the rule using no IV begins to conflate the treatment effects and performance declines.

B.10 Additional Simulations: Overall Population

We conduct additional simulations under the set (B) of assumptions. The settings from the original simulations in Section 3.3.2 are unchanged, except to include sample sizes of 250 and 1000, and consider rules using the linear contrast, $c_{-1} = -1$, $c_0 = 0$, $c_1 = 1$, and the quadratic contrast $c_{-1} = -1/2$, $c_0 = 1$, $c_1 = -1/2$. Figure B.8 presents the misclassification rates of the different decision rules and Figure B.9 presents the value functions for the different decision rules.

The results mimic those seen in Section 3.3.2, where our argmax rule performs well across all success probability $P(A = 1|X, U, Z)$ Scenarios. This can be seen from the low misclassification rates and larger value functions. Further, as the sample size grows, we see our argmax rule continue to approach the oracle rule in performance across all $P(A = 1|X, U, Z)$ Scenarios. As the conditional average treatment effect estimates are confounded by the unmeasured variable U, we see the rule not using the IV generally

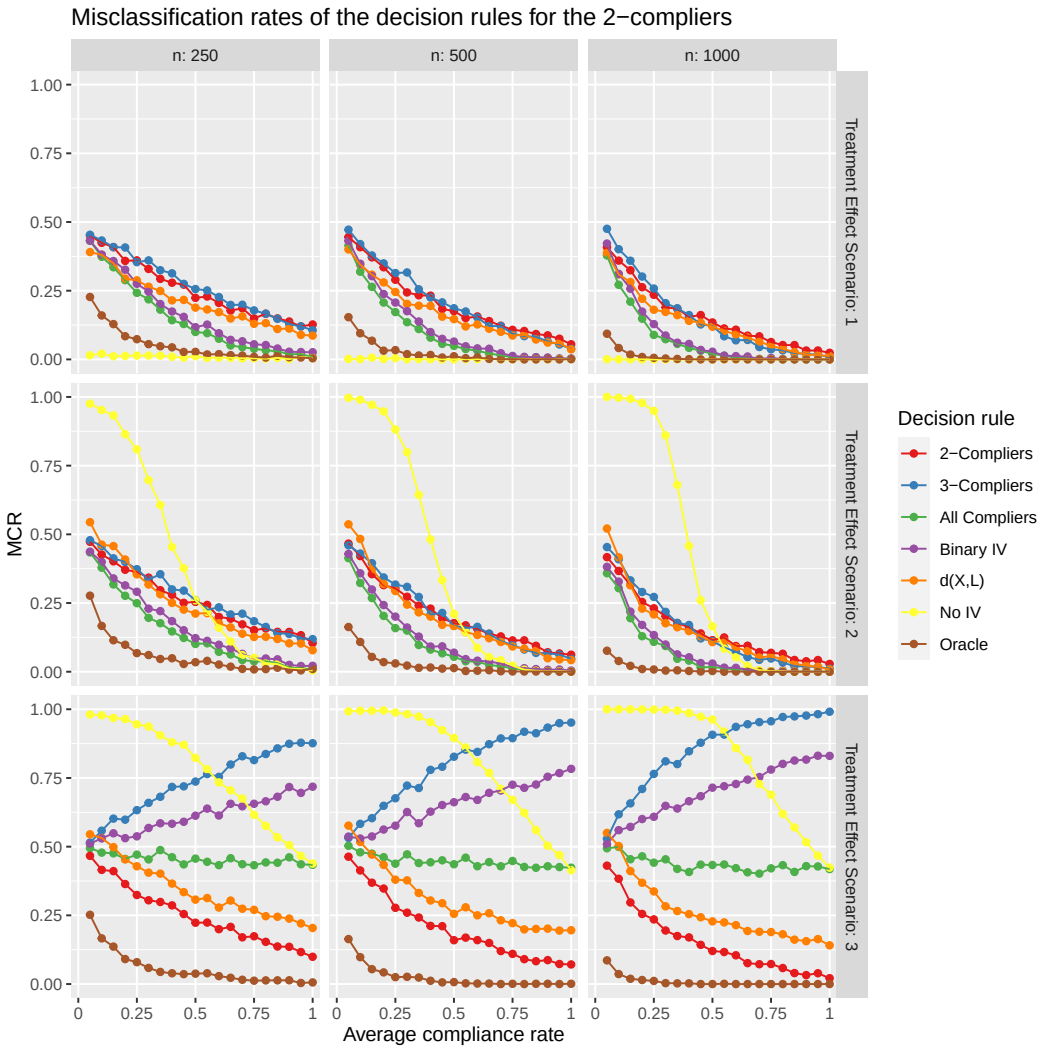


Figure B.4: The average misclassification rates of the different decision rules for the 2-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

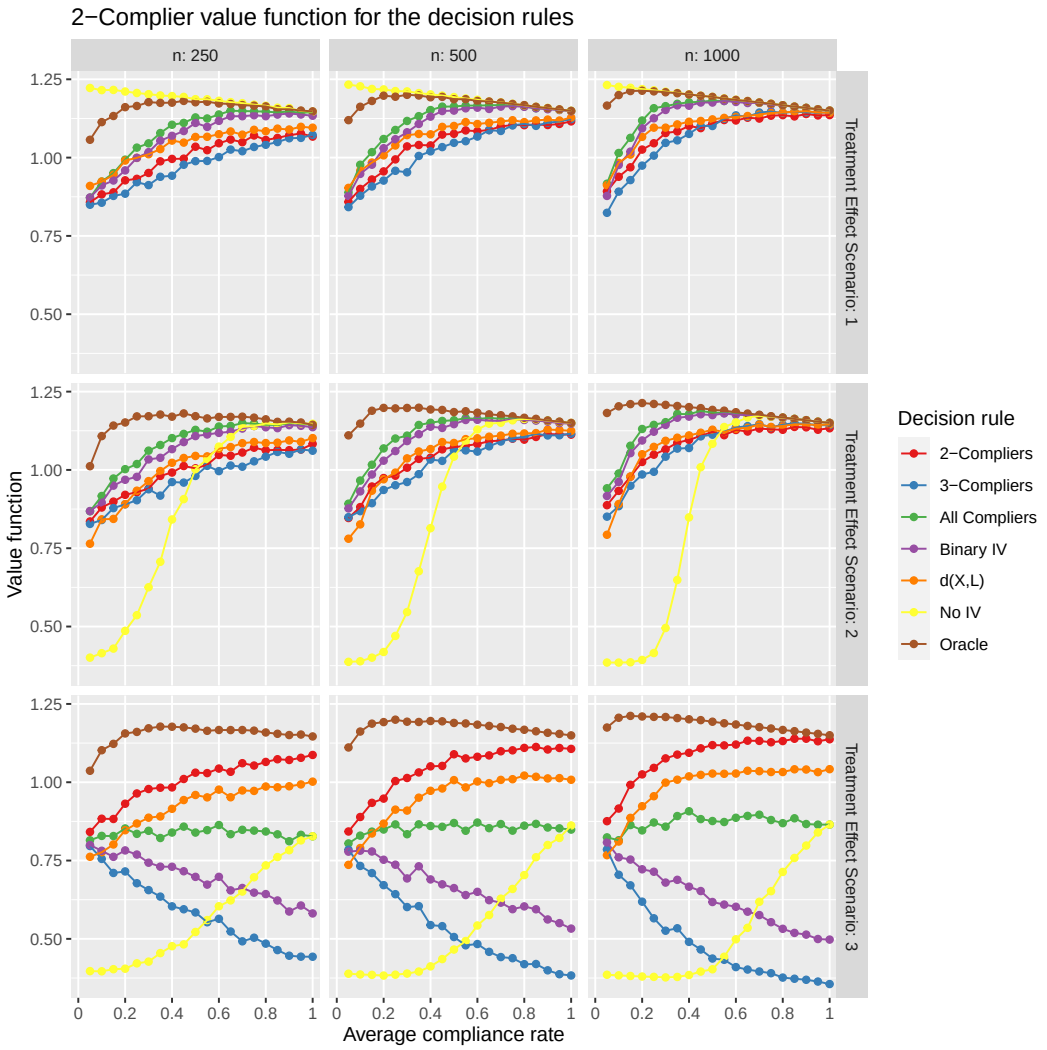


Figure B.5: The average value functions of the different decision rules for the 2-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

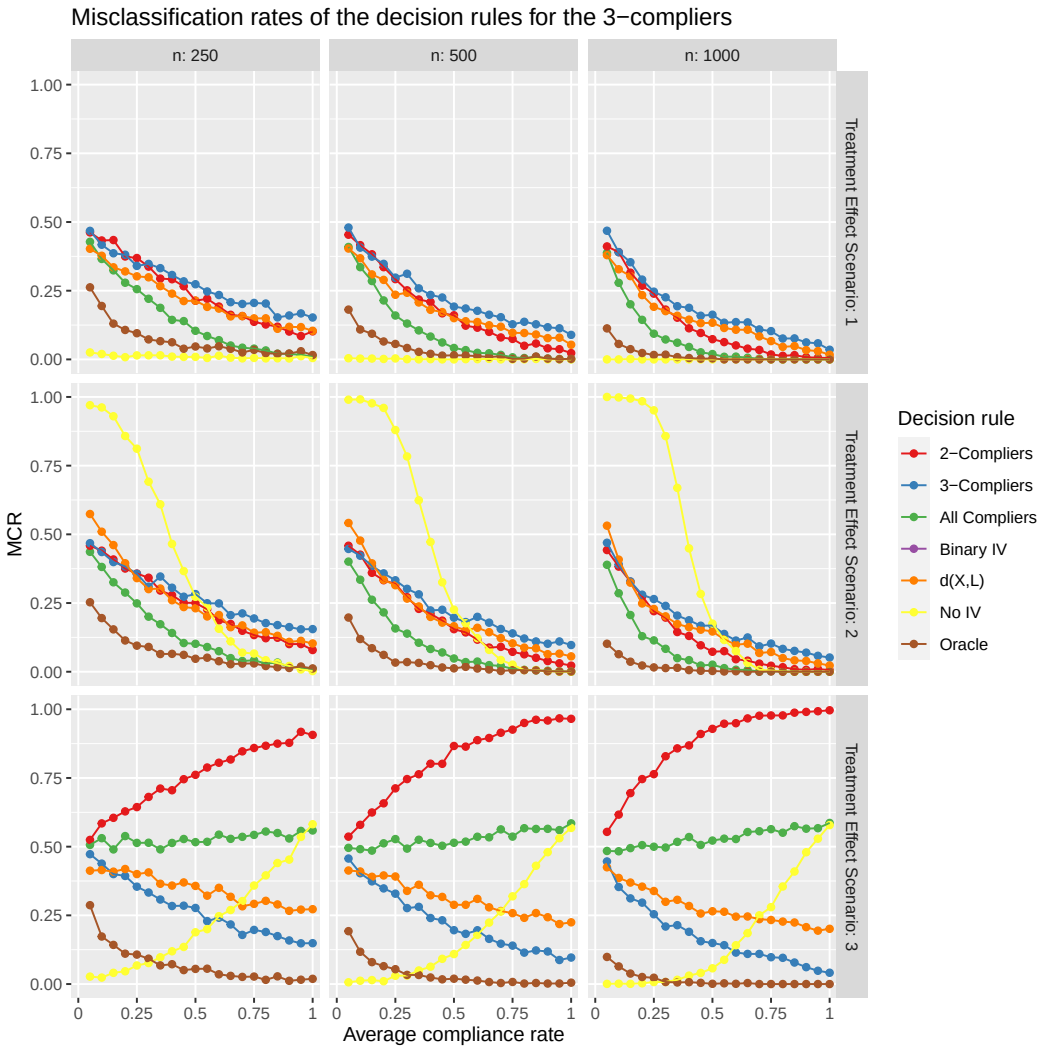


Figure B.6: The average misclassification rates of the different decision rules for the 3-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

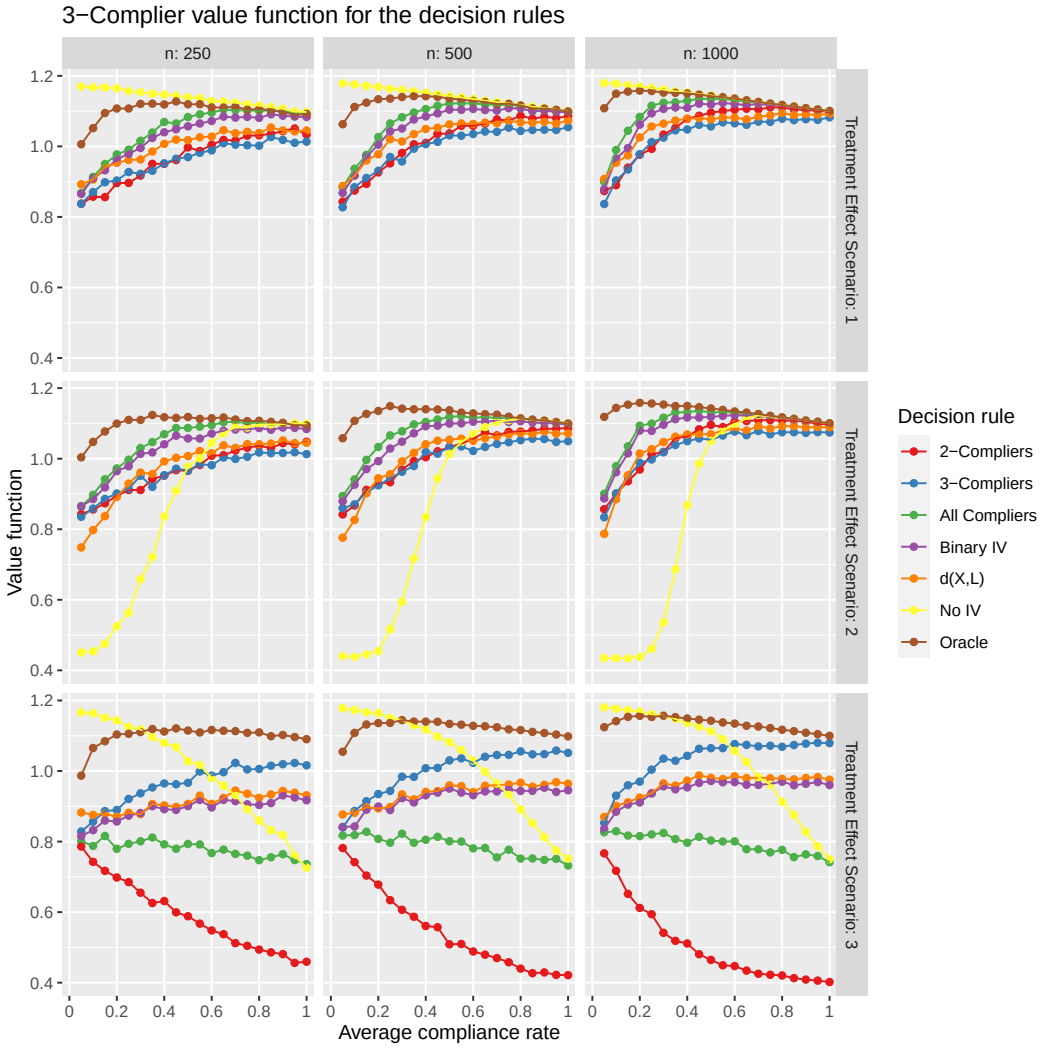


Figure B.7: The average value functions of the different decision rules for the 3-compliers across the three treatment scenarios and for the three training sample sizes. The red line denotes the rule $d_{(2)}(X)$, the blue line denotes the rule $d_{(3)}(X)$, the green line denotes the rule $d_{(all)}(X)$, the purple line denotes the rule using the binary IV, the orange line denotes the rule $\tilde{d}(X, \hat{L})$, the yellow line denotes the rule using no IV, and the brown line denotes the oracle model.

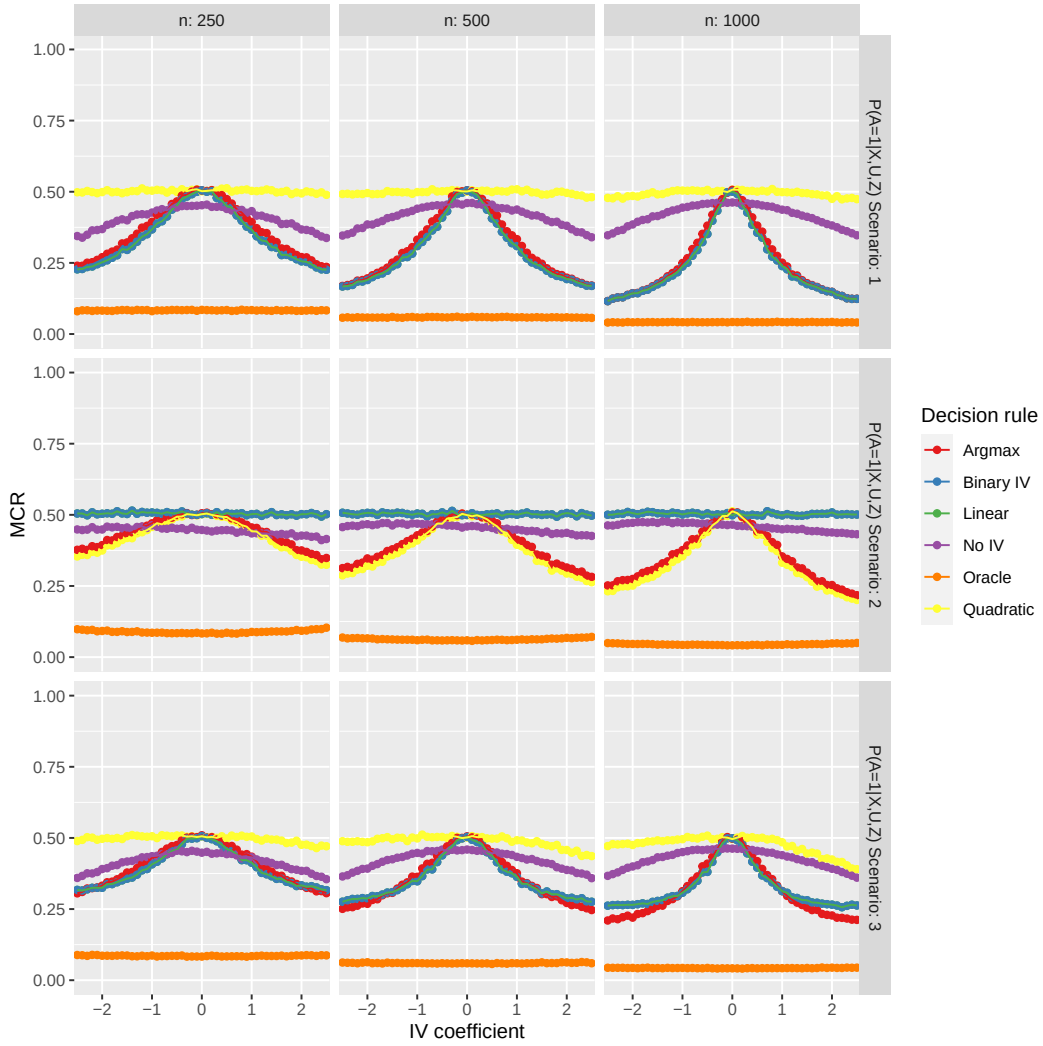


Figure B.8: The average misclassification rates of the different decision rules for the overall population across the three treatment scenarios and for the three training sample sizes. The red line denotes the argmax rule $d_M(X)$, the blue line denotes the rule using the binary IV, the green line denotes the rule using the linear contrast, the purple line denotes the rule using no IV, the orange line denotes the oracle rule, and the yellow line denotes the quadratic rule.

performing worse than the rules using IV, except at IV coefficients near 0, where the rules using IV are subject to weak instrument bias. As the sample size grows, due to the unmeasured confounder, the rule not using the IV does not improve, reflecting the biased estimates. The rule using the binary IV performs well in $P(A = 1|X, U, Z)$ Scenario (1) and (3), but performs very poorly in $P(A = 1|X, U, Z)$ Scenario (2), where the dichotomization of the multilevel IV negates the information the instrument provides on the probability of receiving the treatment. The rule using the linear contrast performs nearly identically to the rule using the binary IV. This is due to the way the multilevel instrument is distributed, coded, and dichotomized (for the binary IV), where the instrument $Z \in \{-1, 0, 1\}$ follows the Categorical distribution with a probability of success of 1/4 for levels $Z = -1$ and $Z = 0$ and 1/2 for level $Z = 1$, and the binary IV is defined as 1 if $Z = 1$ and -1 otherwise. Under these definitions, the instrument has a very similar affect on the probability of receiving the treatment for the multilevel IV with linear contrast and the binary IV. The quadratic contrast only performs well for success probability $P(A = 1|X, U, Z)$ Scenario (2), where the higher and lower level of the IV have a similar effect on the probability of receiving the treatment. It is in this Scenario where the quadratic contrast is most appropriate as it combines the levels which are similar ($Z = -1$ and $Z = 1$) to compare it with the dissimilar level ($Z = 0$). For $P(A = 1|X, U, Z)$ Scenario (3), the performance of the rule using the quadratic contrast is somewhat surprising, as approximately half of the data is generated using the success probability from $P(A = 1|X, U, Z)$ Scenario (2) and so expected to perform better. This is likely due to the fact that the difference between $P(A = 1|X, U, Z = -1)$ and $P(A = 1|X, U, Z = 1)$ in Scenario (1) is greater than the difference between $P(A = 1|X, U, Z = 0)$ and $P(A = 1|X, U, Z = 1)$ in Scenario (2) and so resulting in a slightly weaker instrument for Scenario (2). This can also be seen by comparing the results between $P(A = 1|X, U, Z)$ Scenario (1) and $P(A = 1|X, U, Z)$ Scenario (2) where the performance of the argmax rule and the quadratic rule is slightly weaker in Scenario (2) relative to the argmax rule and linear rule in Scenario (1).

B.11 Additional Analysis: Estimation of Overall Population Value Function Using Single Contrasts

We provide the results of our argmax rule, the rule using the binary IV, and the rule using no IV here when using a single contrast to estimate the value function for the overall

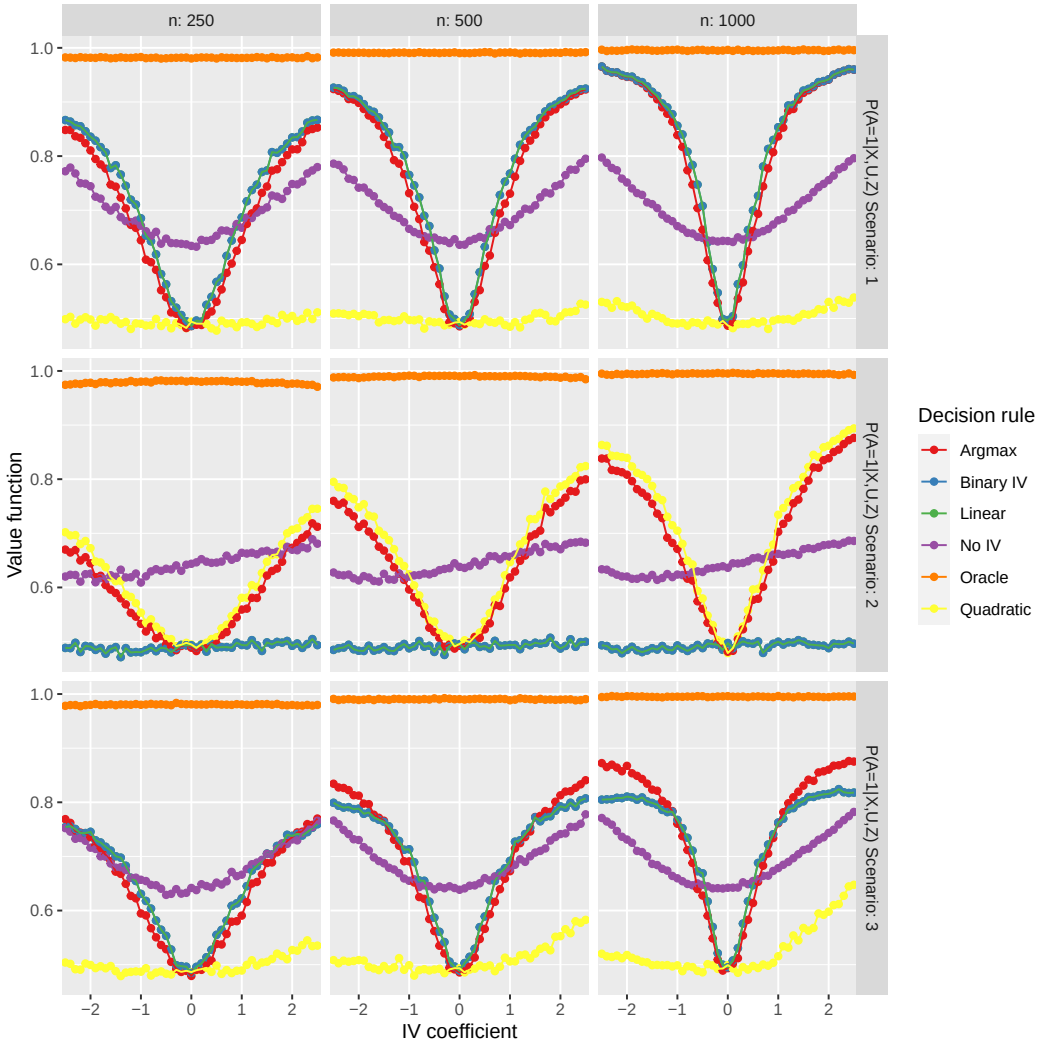


Figure B.9: The average value functions of the different decision rules for the overall population across the three treatment scenarios and for the three training sample sizes. The red line denotes the argmax rule $d_M(X)$, the blue line denotes the rule using the binary IV, the green line denotes the rule using the linear contrast, the purple line denotes the rule using no IV, the orange line denotes the oracle rule, and the yellow line denotes the quadratic rule.

population. The contrasts used to estimate the value function are (c1) comparing the lowest level of the IV with the higher two levels, $c_1 = -1$, $c_2 = c_3 = 1/2$; (c2) comparing the middle level of the IV with the highest and lowest levels, $c_1 = -1/2$, $c_2 = 1$, $c_3 = -1/2$; (c3) comparing the lower two levels of the IV with the highest level, $c_1 = c_2 = -1/2$, $c_3 = 1$; (c4) comparing the lowest and middle levels of the IV, $c_1 = -1$, $c_2 = 1$, $c_3 = 0$; (c5) comparing the lowest and highest levels of the IV, $c_1 = -1$, $c_2 = 0$, $c_3 = 1$; and (c6) comparing the middle and highest levels of the IV, $c_1 = 0$, $c_2 = -1$, $c_3 = 1$. Other than to use a single contrast to estimate the value function, the analysis was conducted as described in Section 3.4.3.

We present the mean (and standard error) of the difference in value functions across the 200 random splits for the different methods, under the assumptions (B1)-(B4) for the overall population, where the value functions are estimated using one of the contrasts (c1)-(c6) in Table B.1. We see that our proposed method performs best for almost all contrasts, except for contrast (c2), as seen by either most reducing the amount of all-cause mortality and stroke within 30-days or causing the least amount of harm as compared to the treatment regime of assigning CEA to all patients. In particular, our proposed argmax rule is estimated to provide the most benefit under contrasts (c1), (c3), (c5), and (c6), and estimated to provide the least harm under contrast (c4). It is for contrasts (c2) and (c4) that all rules are estimated to commit harm relative to assigning CEA to all patients. It is also for these contrasts that the standard errors of the estimated value functions are largest. Due to the large standard errors and our understanding of how the instrument influences the treatment, we believe that contrasts (c2) and (c4) may be artificially imposing a weak instrument bias by negating some of the IV's effect on the probability of receiving the treatment. Because the IV is the center's preference for the CEA procedure, we know that an increasing level of the IV results in an increased probability to receive CEA. A contrast treating the lowest level and the highest level as equal, as in contrast (c2), is likely to result in poor estimation of the value function. This notion is reflected in the larger standard error. Further, ignoring the highest level of the IV which results in the largest probability of receiving CEA, appears to lose too much of the information of how the IV affects the treatment. This is seen in the relatively large standard errors of the estimated value functions under contrast (c4) relative to the other contrasts. As contrasts (c1), (c3), (c5), and (c6) are more stable and contrasts (c2) and (c4) are likely subject to weak instrument bias, we believe that contrasts (c1), (c3), (c5), and (c6) are the better representations of the value function for the given decision rules.

Contrast	Method	$V_{(B)}(d) - V_{(B)}(CEA)$
(c1)	No IV	0.0037 (0.0115)
	Binary IV	-0.0089 (0.0181)
	Argmax	**-0.0124** (0.0163)
(c2)	No IV	**0.0139** (0.0230)
	Binary IV	0.0361 (0.0346)
	Argmax	0.0247 (0.0342)
(c3)	No IV	-0.0027 (0.0077)
	Binary IV	0.0020 (0.0134)
	Argmax	**-0.0060** (0.0127)
(c4)	No IV	0.0659 (0.1989)
	Binary IV	0.1071 (0.2621)
	Argmax	**0.0657** (0.2579)
(c5)	No IV	-0.0005 (0.0068)
	Binary IV	-0.0061 (0.0102)
	Argmax	**-0.0171** (0.0097)
(c6)	No IV	0.0000 (0.0097)
	Binary IV	-0.0029 (0.0156)
	Argmax	**-0.0191** (0.0148)

Table B.1: Mean (and standard error) of the difference between the estimated overall population value function for the different decision methods and the estimated overall population value function for the decision rule of assigning every patient CEA across the 200 random splits of the carotid-artery data. The value functions were estimated using the corresponding contrast.

In comparison to the difference in value functions estimated using the argmax contrasts $M(X)$ in Table 3.3, the estimates of the difference in value functions using single contrasts in Table B.1 are more modest for the argmax rule. As we assume (B4) no additive U-Z interaction is satisfied for every contrast, we expect these estimates to be somewhat similar so long as no weak instrument bias is artificially introduced. There are two possibilities for this discrepancy, either the estimation of the value function using the argmax contrasts is biased or it is a more efficient estimate. It is possible the estimates of the value function are biased due to using contrasts that are correlated with the expected outcome, however, the results in Appendix B.11 imply this may not be the case. It is also possible that the dynamic use of contrasts for certain covariate values allows for gains in efficiency, as it uses certain contrasts for different values of the observed covariates. As there may be an interaction with the IV and a covariate resulting in different relationships of the IV on the probability of receiving the treatment, some contrasts may be more appropriate for estimating the value function at certain values of the covariates then others, resulting in an increase in efficiency. Regardless, using either the argmax contrasts or the single contrasts to estimate the value function largely agree in their consensus that the argmax rule leads to the greatest benefit in reducing stroke or death within 30-days, as compared to a rule using a binary IV and a rule not using an IV.

Finally, we note that the the rules using an IV generally outperform the rule not using the IV. For the contrasts that are not expected to be subject to an artificial weak instrument bias, (c1), (c3), (c5), and (c6), the rules using an IV generally outperform the rule not using an IV. It is only for the contrast (c3) that the rule not using an IV is estimated to outperform the rule using a binary IV. This is likely due to an unmeasured confounder biasing the conditional average treatment effect and preventing the rule not using an IV from identifying an optimal ITR.

REFERENCES

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

Ai, Chunrong, and Xiaohong Chen. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71(6):1795–1843.

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens. 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies* 67(3):499–527.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–455.

Athey, Susan, and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey, Susan, and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *arXiv:1504.01132v1 [stat.ML].*

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *The Annals of Statistics* 47(2):1148–1178.

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

Baiocchi, Mike, Dylan S Small, Scott Lorch, and Paul R Rosenbaum. 2010. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* 105(492):1285–1296.

Balke, Alexander, and Judea Pearl. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439):1171–1176.

Bargagli-Stoffi, Falco J, Kristof De-Witte, and Giorgio Gnecco. 2019. Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach. *arXiv:1905.12707 [stat.ME].*

Bargagli-Stoffi, Falco J, and Giorgio Gnecco. 2018. Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In *2018 ieee 5th international conference on data science and advanced analytics (dsaa)*, 1–10. IEEE.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica* 75(6):1613–1669.

Blundell, Richard, and James L Powell. 2003. Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs* 36:312–357.

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone. 1984. *Classification and regression trees.* New York: Chapman and Hall/CRC.

Brookhart, M Alan, and Sebastian Schneeweiss. 2007. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics* 3(1):1–25.

Brott, Thomas G, George Howard, Gary S Roubin, James F Meschia, Ariane Mackey, William Brooks, Wesley S Moore, Michael D Hill, Vito A Mantese, Wayne M Clark, et al. 2016. Long-term results of stenting versus endarterectomy for carotid-artery stenosis. *New England Journal of Medicine* 374(11):1021–1031.

Chen, Guanhua, Donglin Zeng, and Michael R Kosorok. 2016. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association* 111(516): 1509–1521.

Chen, Jingxiang, Haoda Fu, Xuanyao He, Michael R Kosorok, and Yufeng Liu. 2018. Estimating individualized treatment rules for ordinal treatments. *Biometrics* 74(3):924–933.

Chen, Shuai, Lu Tian, Tianxi Cai, and Menggang Yu. 2017. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* 73(4):1199–1209.

Chen, Shuxiao, and Bo Zhang. 2021. Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *arXiv preprint arXiv:2104.07822*.

Chen, Xiaohong, and Demian Pouzo. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80(1): 277–321.

Cheng, Jing, and Dylan S Small. 2006. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(5):815–836.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *National Bureau of Economic Research*.

Cole, J Alexander, Heather Norman, Lisa B Weatherby, and Alexander M Walker. 2006. Drug copayment and adherence in chronic heart failure: effect on cost and outcomes. *Pharmacotherapy* 26(8):1157–1164.

Columbo, Jesse A, Pablo Martinez-Camblor, Todd A MacKenzie, Douglas O Staiger, Ravinder Kang, Philip P Goodney, and A James O'Malley. 2018. Comparing long-term mortality after carotid endarterectomy vs carotid stenting using a novel instrumental variable method for risk adjustment in observational time-to-event data. *JAMA Network* 1(5):1–14.

CPMP, Working Party on Efficacy of Medicinal Products. 1995. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products: note for guidance. *Statistics in Medicine* 14(15):1659–1682. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780141507`.

Cui, Yifan, and Eric Tchetgen Tchetgen. 2020. On a necessary and sufficient identification condition of optimal treatment regimes with an instrumental variable. *arXiv preprint arXiv:2010.03390*.

———. 2021a. Machine intelligence for individualized decision making under a counterfactual world: A rejoinder. *Journal of the American Statistical Association* 116(533): 200–206.

———. 2021b. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association* 116(533): 162–173.

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. Nonparametric instrumental regression. *Econometrica* 79(5):1541–1565.

Ding, Peng. 2017. A paradox from randomization-based causal inference. *Statistical Science* 32(3):331–345.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. The oregon health insurance experiment: evidence from the first year. *The Quarterly Journal of Economics* 127(3):1057–1106.

Fogarty, Colin B. 2018. Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika* 105(4):994–1000.

———. 2020. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association* 115(531): 1518–1530.

Fogarty, Colin B, Kwonsang Lee, Rachel R Kelz, and Luke J Keele. 2021. Biased encouragements and heterogeneous effects in an instrumental variable study of emergency general surgical outcomes. *Journal of the American Statistical Association*.

Frangakis, Constantine E, and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58(1):21–29.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1): 1–22.

Giles, Kristina A, Allen D Hamdan, Frank B Pomposelli, Mark C Wyers, and Marc L Schermerhorn. 2010. Stroke and death after carotid endarterectomy and carotid artery stenting with and without high risk criteria. *Journal of Vascular Surgery* 52(6):1497–1504.

Goyal, Neera, José R Zubizarreta, Dylan S Small, and Scott A Lorch. 2013. Length of stay and readmission among late preterm infants: an instrumental variable approach. *Hospital Pediatrics* 3(1):7–15.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. 2020. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis* 15(3):965 – 1056.

Hall, Peter, and Joel L Horowitz. 2005. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics* 33(6):2904–2929.

Halm, Ethan A, Clara Lee, and Mark R Chassin. 2002. Is volume related to outcome in health care? a systematic review and methodologic critique of the literature. *Annals of Internal Medicine* 137(6):511–520.

Hernán, Miguel A, and James M Robins. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 360–372.

Hilbert, Martin, and Priscila López. 2011. The world's technological capacity to store, communicate, and compute information. *Science* 332(6025):60–65.

Hill, Jennifer L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–240.

Hodges, Joseph L, and Erich L Lehmann. 1963. Estimates of location based on rank tests. *The Annals of Mathematical Statistics* 34(2):598–611.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–960.

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

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

Huling, Jared D, Menggang Yu, and A James O'Malley. 2019. Instrumental variable based estimation under the semiparametric accelerated failure time model. *Biometrics* 75(2): 516–527.

Imbens, Guido W. 2010. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature* 48(2):399–423.

Imbens, Guido W, and Joshua D Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 61(2):467–476.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2019. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, 2281–2290. PMLR.

Kallus, Nathan, and Angela Zhou. 2019. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*.

Kang, Hyunseung, Benno Kreuels, Ohene Adjei, Ralf Krumkamp, Jürgen May, and Dylan S Small. 2013. The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International Journal of Epidemiology* 42(5):1390–1398.

Kang, Hyunseung, Benno Kreuels, Jürgen May, and Dylan S Small. 2016a. Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *The Annals of Applied Statistics* 10(1):335–364.

Kang, Hyunseung, Laura Peck, and Luke Keele. 2018. Inference for instrumental variables: a randomization inference approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(4):1231–1254.

Kang, Hyunseung, Anru Zhang, T Tony Cai, and Dylan S Small. 2016b. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* 111(513):132–144.

Kennedy, Edward H, Scott Lorch, and Dylan S Small. 2019. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(1):121–143.

Lee, Kwonsang, Falco J Bargagli-Stoffi, and Francesca Dominici. 2021a. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*.

Lee, Kwonsang, Dylan S. Small, and Francesca Dominici. 2021b. Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *Journal*

of the American Statistical Association 116(534):569–580. https://doi.org/10.1080/01621459.2020.1870476.

Lee, Kwonsang, Dylan S Small, Jesse Y Hsu, Jeffrey H Silber, and Paul R Rosenbaum. 2018a. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(2):535–546.

Lee, Kwonsang, Dylan S Small, and Paul R Rosenbaum. 2018b. A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* 74(4): 1161–1170.

Liang, Muxuan, and Menggang Yu. 2020. A semiparametric approach to model effect modification. *Journal of the American Statistical Association* 1–13.

Liaw, Andy, and Matthew Wiener. 2002. Classification and regression by randomforest. *R News* 2(3):18–22.

Lorch, Scott A, Michael Baiocchi, Corinne E Ahlberg, and Dylan S Small. 2012. The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics* 130(2):270–278.

Lou, Zhilan, Jun Shao, and Menggang Yu. 2018. Optimal treatment assignment to maximize expected outcome with multiple treatments. *Biometrics* 74(2):506–516.

Luft, Harold S, John P Bunker, and Alain C Enthoven. 1979. Should operations be regionalized? the empirical relation between surgical volume and mortality. *New England Journal of Medicine* 301(25):1364–1369.

Malkin, Jesse D, Michael S Broder, and Emmett Keeler. 2000. Do longer postpartum stays reduce newborn readmissions? analysis using instrumental variables. *Health Services Research* 35(5 Pt 2):1071–1091.

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

Martínez-Camblor, Pablo, Todd Mackenzie, Douglas O Staiger, Philip P Goodney, and A James O'Malley. 2019a. Adjusting for bias introduced by instrumental variable estimation in the cox proportional hazards model. *Biostatistics* 20(1):80–96.

Martínez-Camblor, Pablo, Todd A MacKenzie, Douglas O Staiger, Phillip P Goodney, and A James O'Malley. 2019b. An instrumental variable procedure for estimating cox models with non-proportional hazards in the presence of unmeasured confounding. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68(4):985–1005.

Mcclellan, Mark, Barbara J Mcneil, and Joseph P Newhouse. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? analysis using instrumental variables. *Journal of the American Medical Association* 272(11):859–866.

McPhee, James T, Andres Schanzer, Louis M Messina, and Mohammad H Eslami. 2008. Carotid artery stenting has increased rates of postprocedure stroke, death, and resource utilization than does carotid endarterectomy in the united states, 2005. *Journal of Vascular Surgery* 48(6):1442–1450.

Moodie, Erica EM, Nema Dean, and Yue Ru Sun. 2014. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences* 6(2):223–243.

Newey, Whitney K, and James L Powell. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71(5):1565–1578.

Neyman, Jersey. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes (excerpts reprinted and translated to english, 1990). *Statistical Science* 5:463–472.

Nolan, Brian W, Randall R De Martino, Philip P Goodney, Andres Schanzer, David H Stone, David Butzel, Christopher J Kwolek, Jack L Cronenwett, Vascular Study Group of New England, et al. 2012. Comparison of carotid endarterectomy and stenting in real world practice using a regional quality improvement registry. *Journal of Vascular Surgery* 56(4):990–996.

Okui, Ryo, Dylan S Small, Zhiqiang Tan, and James M Robins. 2012. Doubly robust instrumental variable regression. *Statistica Sinica* 173–205.

O'Malley, A James, Philip Cotterill, Marc L Schermerhorn, and Bruce E Landon. 2011. Improving observational study estimates of treatment effects using joint modeling of selection effects and outcomes: the case of aaa repair. *Medical Care* 49(12):1126–1132.

Park, Chan, and Hyunseung Kang. 2020. A groupwise approach for inferring heterogeneous treatment effects in causal inference. *arXiv preprint arXiv:1908.04427v2*.

Pu, Hongming, and Bo Zhang. 2020. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *arXiv e-prints* arXiv–2002.

Qian, Min, and Susan A Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39(2):1180–1210.

Qiu, Hongxiang, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. 2021. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association* 116(533):174–191.

Rosenbaum, Paul R. 2002a. Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 17(3):286–327.

———. 2002b. [covariance adjustment in randomized experiments and observational studies]: Rejoinder. *Statistical Science* 17(3):321–327.

———. 2010. *Design of observational studies*. New York: Springer.

———. 2020. Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application* 7:143–176.

Rosenfield, Kenneth, Jon S Matsumura, Seemant Chaturvedi, Tom Riles, Gary M Ansel, D Chris Metzger, Lawrence Wechsler, Michael R Jaff, and William Gray. 2016. Randomized trial of stent versus surgery for asymptomatic carotid stenosis. *New England Journal of Medicine* 374(11):1011–1020.

Rothwell, Peter M. 2005. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* 365(9454):176–186.

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

———. 1980. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association* 75(371):591–593.

———. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2(3-4): 169–188.

Staiger, Douglas O, and James H Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–586.

Stallones, Reuel A. 1987. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine* 16(2):183–194.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20(4):518–529.

Stuart, Elizabeth A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1):1–21.

Su, Liangjun, Irina Murtazashvili, and Aman Ullah. 2013. Local linear gmm estimation of functional coefficient iv models with an application to estimating the rate of return to schooling. *Journal of Business & Economic Statistics* 31(2):184–207.

Su, Xiaogang, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10: 141–158.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Swanson, Sonja A, and Miguel A Hernán. 2013. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 24(3):370–374.

———. 2014. Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. *Statistical Science* 29(3):371–374.

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

Therneau, Terry, Beth Atkinson, and Brian Ripley. 2015. *Package 'rpart'*. R package version 4.1-15. https://cran.r-project.org/package=rpart.

Tian, Lu, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. 2014. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* 109(508):1517–1532.

Wager, Stefan, and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523): 1228–1242.

Wald, Abraham. 1940. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics* 11(3):284–300.

Wang, Fen Wei, Dennis Esterbrooks, Yong-Fang Kuo, Aryan Mooss, Syed M Mohiuddin, and Barry F Uretsky. 2011. Outcomes after carotid artery stenting and endarterectomy in the medicare population. *Stroke* 42(7):2019–2025.

Wang, Linbo, and Eric Tchetgen Tchetgen. 2018. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 80(3):531.

Wang, Sheng, and Hyunseung Kang. 2021. Weak-instrument robust tests in two-sample summary-data mendelian randomization. *arXiv preprint arXiv:1909.06950*.

Wang, Tong, and Cynthia Rudin. 2021. Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*.

Wang, Zengri, and Thomas A Louis. 2003. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* 90(4): 765–775.

Xu, Yaoyao, Menggang Yu, Yingqi Zhao, Quefeng Li, Sijian Wang, and Jun Shao. 2015. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* 71(3):645–653.

Ye, Ting, Jun Shao, and Hyunseung Kang. 2021. Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of Statistics* 49(4):2079–2100.

Yu, Ruoqi. 2019. *bigmatch: Making optimal matching size-scalable using optimal calipers*. R package version 0.6.1. https://CRAN.R-project.org/package=bigmatch.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, and Herman A. Tyroler. 1991. Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials. *JAMA* 266(1):93–98. `https://jamanetwork.com/journals/jama/articlepdf/386387/jama_266_1_038.pdf`.

Zhang, Baqun, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. 2012. Estimating optimal treatment regimes from a classification perspective. *Stat* 1(1):103–114.

Zhao, Yingqi, Donglin Zeng, A John Rush, and Michael R Kosorok. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107(499):1106–1118.

Zhao, Yufan, Michael R Kosorok, and Donglin Zeng. 2009. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine* 28(26):3294–3315.

Zhou, Xin, and Michael R Kosorok. 2017. Augmented outcome-weighted learning for optimal treatment regimes. *arXiv preprint arXiv:1711.10654*.