

---

# Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions

---

Zeshan Hussain\*  
MIT

Ming-Chieh Shih\*  
National Dong Hwa University

Michael Oberst  
MIT

Ilker Demirel  
MIT

David Sontag  
MIT

## Abstract

*Randomized Controlled Trials (RCT)s* are relied upon to assess new treatments, but suffer from limited power to guide personalized treatment decisions. On the other hand, observational (i.e., non-experimental) studies have large and diverse populations, but are prone to various biases (e.g. residual confounding). To safely leverage the strengths of observational studies, we focus on the problem of *falsification*, whereby RCTs are used to validate causal effect estimates learned from observational data. In particular, we show that, given data from both an RCT and an observational study, assumptions on internal and external validity have an observable, testable implication in the form of a set of *Conditional Moment Restrictions (CMRs)*. Further, we show that expressing these CMRs with respect to the causal effect, or “causal contrast”, as opposed to individual counterfactual means, provides a more reliable falsification test. In addition to giving guarantees on the asymptotic properties of our test, we demonstrate superior power and type I error of our approach on semi-synthetic and real world datasets. Our approach is interpretable, allowing a practitioner to visualize which subgroups in the population lead to falsification of an observational study.

erature in building estimators of heterogeneous treatment effects from observational data, particularly using modern machine learning methods (Wager and Athey, 2018; Künzel et al., 2019; Semenova and Chernozhukov, 2021). However, observational studies may lack *internal validity* in that estimates of causal effects (in the observational population) may be biased or inconsistent, e.g., due to unobserved differences between the treatment and control groups, such as in the setting of unobserved confounding. On the other hand, observational studies are representative of more diverse populations, leading to more plausible *external validity*, i.e., ability to generalize estimates across wider populations. In contrast, *Randomized Controlled Trials (RCTs)* have strong internal validity, assuming sound design (e.g. a prospective trial, a-priori definition of hypotheses to be tested) and appropriate randomization. However, RCTs often have restrictive inclusion criteria, which can call their external validity into question (Degtiar and Rose, 2021), and are of limited size, limiting their ability to detect differences in treatment effect for specific sub-populations, or detect differences in adverse event rates (if e.g., the adverse event is rare) (Tsang et al., 2009; Ali et al., 2018; Phillips et al., 2019). Intuitively, we would like to leverage observational data for estimating treatment effects that cannot be reliably estimated using RCTs, whether due to a lack of statistical power or a lack of patient diversity in the RCT. At the same time, we would like to take advantage of the strong internal validity of the RCT to increase confidence in our observational estimates.

## 1 INTRODUCTION

Observational studies, prevalent in healthcare, economics, and other fields, are an important source of real-world data used to derive granular treatment effect estimates (Dagan et al., 2021; Hernán et al., 2008; Imbens and Wooldridge, 2009; Xie et al., 2020). Indeed, there has been a rich lit-

With these considerations in mind, we study the problem of using limited RCT data to “falsify” assumptions of internal and external validity for observational studies. Our method can be applied even when the RCT data does not cover the entire observational population, and hence cannot be used on its own to estimate causal effects. Assuming that the RCT has internal validity, we show that assumptions of internal/external validity of the observational study have a testable implication in the form of a set of conditional moment restrictions (CMRs). We propose a *falsification algorithm* that tests whether or not these CMRs hold, thereby

---

\*equal contribution, order determined alphabetically.

Proceedings of the 26<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

providing an opportunity to reject these assumptions when they fail to hold. This allows us to take advantage of approaches developed in the econometrics literature for testing CMRs, and we use a Maximum Moment Restriction (MMR)-based test (Muandet et al., 2020) for this purpose.

Compared to prior work, the benefits of our approach are two-fold. First, we implicitly check across all subpopulations of covariates  $X$  for disagreement between the conditional average treatment effect (CATE) functions as estimated in the RCT versus in the observational study. Second, as an additional benefit, our approach provides an explanation for rejection, in the form of a “witness function”, which describes subpopulations where these estimates diverge.

Importantly, in constructing the test for our problem, we use the insight that not all differences between observational and experimental distributions matter. For instance, there may be differences in unobserved baseline risk factors, which cause estimates of individual potential outcomes to differ, but do not impact the causal effect (the difference between control and treated outcomes). A naive MMR-based test would asymptotically reject in this scenario, but we demonstrate that, by careful construction of the MMR test statistic, we can avoid this failure case.

Our method can be compared to prior approaches to falsification of observational studies. One such approach is to check for a statistically significant difference between estimates of the average treatment effect (ATE) from the RCT and from the observational study (Franklin et al., 2021; Dagan et al., 2021; Baden et al., 2020). Unfortunately, this approach can lead to false negatives, e.g., if the ATE from the observational study replicates the RCT ATE, despite biases on finer-grained subpopulations. Furthermore, even if an observational study is correctly “rejected”, the approach does not provide an **explanation** for why the observational study was rejected, which is an important practical consideration for both statisticians and policymakers. Another approach is to compare subgroup-level effects instead of the ATE. Hussain et al. (2022) adopt this approach in the context of testing (multiple) observational estimates against RCT estimates, essentially testing for differences in group-wise treatment effects. However, this approach requires correction for multiple hypothesis testing across subgroups, and a-priori specification of these subgroups, which can limit its ability to uncover areas of disagreement.

**Contributions:** We have the following desiderata for our falsification algorithm: (i) rejecting observational studies when their underlying causal assumptions fail (*high power*), (ii) accepting in cases where these casual assumptions hold (*controlled type I error*), and (iii) providing an explanation of why an observational study is rejected. With these desiderata in mind, our main contributions are as follows: (i) First, we show how to convert causal assumptions on internal and external validity into a set of CMRs, viola-

tions of which can be detected using observational and RCT data using existing techniques with theoretical guarantees. (ii) Second, we demonstrate that our construction of these CMRs avoids a potential failure mode: rejecting observational studies due to differences in unobserved covariates that influence baseline outcomes, but not treatment effects. (iii) Third, on semi-synthetic and real-world datasets, we show favorable performance of our method with respect to power and type I error, and showcase its ability to produce informative *explanations* of rejections.

## 2 SETUP AND MOTIVATING EXAMPLES

### 2.1 Notation & Assumptions

Let  $Y \in \mathcal{Y}$  be the outcome of interest, and  $A \in \{0, 1\}$  denote a binary treatment variable. We let  $Y_a$  denote the potential outcome under treatment  $A = a$ , and we use  $X \in \mathcal{X}$  to denote the full set of covariates. Note that, in our development, we operate in the setting where there is a single observational study and RCT. We use an indicator variable,  $S = \{0, 1\}$ , where  $S = 1$  denotes data from the observational study and  $S = 0$  from the RCT.

To further characterize the observational study and RCT, we let  $\mathcal{I}_0$  and  $\mathcal{I}_1$  be the observed indices for the RCT and observational study, respectively. Furthermore, we let  $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$  be the total set of observed indices. We use  $|\mathcal{I}|$  to denote the cardinality of a set, and let  $|\mathcal{I}_0| = n_0$ ,  $|\mathcal{I}_1| = n_1$ , and  $|\mathcal{I}| = n$ . Finally,  $\mathbb{E}[\cdot]$  and  $\mathbb{P}[\cdot]$  are expectations and probabilities taken with respect to the joint distribution  $\mathbb{P}(Y, A, X, S)$  of the observational study and RCT.

Our goal is to discover violations of causal assumptions that underlie the validity of conditional average treatment effect (CATE) estimates derived from the observational study. To that end, we first state these assumptions formally.

**Assumption 2.1 (Internal Validity of Observational Data).** We assume the following in the observational study:

- *Ignorability* —  $Y_a \perp\!\!\!\perp A \mid X, (S = 1), \forall a \in \{0, 1\}$ .
- *Consistency* —  $A = a, S = 1 \implies Y_a = Y, \forall a \in \{0, 1\}$ .
- *Positivity of Treatment* —  $\mathbb{P}(X = x, S = 1) > 0 \implies \mathbb{P}(A = a \mid X = x, S = 1) > 0, \forall a \in \{0, 1\}$  and  $\forall x \in \mathcal{X}$ .

Assumption 2.1 gives a standard set of assumptions under which the CATE conditioned on  $X$ ,  $\mathbb{E}[Y_1 - Y_0 \mid X = x, S = 1]$  can be identified. However, this assumption is not testable in isolation. In order to compare observational estimates with those of the RCT to discover flaws, we will first need to assume that the RCT itself provides valid estimates.

**Assumption 2.2 (Internal Validity of RCT).** We assume the following in the RCT:

- *Ignorability* —  $Y_a \perp\!\!\!\perp A \mid X, (S = 0), \forall a \in \{0, 1\}$ .
- *Consistency* —  $A = a, S = 0 \implies Y_a = Y, \forall a \in \{0, 1\}$ .

- *Fixed probability of assignment* —  $P(A = 1|X = x, S = 0) = p$ , for some  $p \in (0, 1)$ ,  $\forall x \in \mathcal{X}$ .

Assumption 2.2 is a generally defensible (and standard) set of assumptions on the validity of the RCT. However, even if both Assumptions 2.1 and 2.2 hold, the corresponding CATE functions are not necessarily comparable. For instance, there may be unmeasured effect modifiers that have different distributions between the RCT and observational study. Under the following additional assumption, the CATE in the RCT (i.e.,  $\mathbb{E}[Y_1 - Y_0 | X = x, S = 0]$ ) can be identified using observational data.

**Assumption 2.3** (*External Validity: Observational Study to RCT Transportability of CATE*). We assume the following:

- *Mean Exchangeability of Contrast* —  $\mathbb{E}[Y_1 - Y_0|X = x] = \mathbb{E}[Y_1 - Y_0|X = x, S = s]$ ,  $\forall x \in \mathcal{X}$  and  $\forall s \in \{0, 1\}$ .
- *Positivity of Selection* —  $\mathbb{P}(X = x|S = 0) > 0 \implies \mathbb{P}(X = x|S = 1) > 0$ ,  $\forall x \in \mathcal{X}$ .

The first part of this assumption is sometimes referred to as “generalizability in effect measure” (Dahabreh et al., 2019) or “conditional exchangeability in measure” (Dahabreh et al., 2020). This assumption is weaker than (and implied by) transportability of counterfactual means (e.g.,  $\mathbb{E}[Y_a | X, S] = \mathbb{E}[Y_a | X]$ )<sup>1</sup>. It is simple to show that this assumption (along with our other assumptions) is sufficient to identify the CATE in the RCT population using the observational distribution alone.

**Proposition 2.1.** *Under Assumptions 2.1 and 2.3, the CATE of the RCT given  $X$ ,  $\mathbb{E}[Y_1 - Y_0|X, S = 0]$ , is identifiable in the observational data by*

$$\mathbb{E}[Y | X, A = 1, S = 1] - \mathbb{E}[Y | X, A = 0, S = 1] \quad (1)$$

Proposition 2.1 follows from substantially the same arguments used by Dahabreh et al. (2019) for identification of average treatment effects under similar assumptions, but we include a short proof in Appendix A for completeness, alongside all other proofs for this paper.

*Remark 2.1.* As we demonstrate later on, our statistical test does not distinguish between violations of Assumption 2.1 or Assumption 2.3. However, violation of either assumption is a meaningful finding when considering the credibility of causal effects learned from observational data. For instance, even if the observational study is free of unmeasured confounding (i.e., Assumption 2.1 holds), there may exist unmeasured effect modifiers whose distributions differ substantially across populations, leading to a violation of Assumption 2.3. In other words, if the true CATE function varies substantially for individuals with the same covariates  $X$  across the observational and RCT populations, then it may not reliably generalize to future patients.

<sup>1</sup>Equality relations including random variables are to be understood as “almost sure” (a.s.) relations throughout the manuscript.

## 2.2 Motivation: Testing for Differences in Causal Contrasts, rather than Counterfactual Means

When it is possible to identify a causal effect from an observational study, we would prefer to avoid rejecting that study unnecessarily. This motivates Assumption 2.3, which holds even in the scenarios where counterfactual means in the RCT (e.g., the expected outcome under treatment  $\mathbb{E}[Y_1 | X, S = 0]$ ) are **not** identifiable from observational data, but where the causal contrast is identified.

This assumption is central to our testing methodology, as we test for a null hypothesis that is satisfied under Assumption 2.3, even when counterfactual means are not transportable. We build intuition for this assumption in two ways. First, we give a structural causal model that formalizes a sufficient condition for this assumption to hold. Second, we give concrete examples of where this assumption appears to (approximately) hold in practice.

**Example 2.1.** Let  $U$  be a set of variables that are unobserved in both the RCT and observational data. Suppose  $Y$  is generated according to the following structural equation, with binary treatment  $A$  and observed covariates  $X$

$$Y = g(X, U) + \tau(X) \cdot A + \epsilon_0, \quad (2)$$

where  $\epsilon_0$  is an independent mean-zero random variable ( $\mathbb{E}[\epsilon_0] = 0$  and  $\epsilon_0 \perp X, U, A, S$ ) and where  $P(U | X, S = 1) \neq P(U | X, S = 0)$ .

In Example 2.1,  $Y_0 = g(X, U) + \epsilon_0$  is influenced by both  $X$  and a set of unobserved baseline characteristics  $U$ . As a result, the conditional counterfactual mean  $\mathbb{E}[Y_0 | X, S] = \mathbb{E}[g(X, U) | X, S]$  will generally differ across studies, due to the fact that the distribution of  $U$  varies across studies. The conditional average treatment effect, on the other hand, is independent of  $U$  and  $S$ , as  $\mathbb{E}[Y_1 - Y_0 | X] = \tau(X)$ . This quantity is purely a function of  $X$ , satisfying our assumption that the CATE does not depend on  $S$ .<sup>2</sup>

This scenario is plausible in real-world settings, where the treatment effect is a function of a subset of variables that influence the outcome  $Y$ . For a real-world example, consider the SPRINT Trial (SPRINT Research Group, 2015), which studies the impact of intensive blood pressure control ( $A$ ) on a composite outcome ( $Y$ ) that includes heart failure and death. Here, previous chronic kidney disease (CKD) is a variable, like  $U$ , that has a substantial impact on the outcome  $Y_0$  under no treatment (as reported in Figure 4 of SPRINT Research Group (2015)), but does not have a (statistically) significant influence on the treatment effect itself (i.e.,  $\tau(X)$  in the example above). We discuss this example and other examples of real-world motivation in more detail in Appendix B.

<sup>2</sup>The constant treatment effect for individuals with the same  $X$  is not necessary, and merely helps simplify notation. One could make a similar observation with  $\tau(X, \epsilon_\tau)$  for an additional noise variable  $\epsilon_\tau$  that is independent of  $U, S$ .

### 3 MMR-based FALSIFICATION TESTS

Next, we observe that Assumptions 2.1 to 2.3 have observable implications on the joint RCT and observational data in the form of a (set of) conditional moment restrictions. As a result, if these restrictions fail to hold, then this implies a violation of the underlying causal assumptions. This suggests a hypothesis-testing approach for detecting violations, which we develop in this section. Notably, the resulting hypothesis test looks for differences between the CATE functions estimated from the RCT and observational studies, but does not test for equality of conditional potential outcomes themselves. This is motivated from our prior discussion, that conditional means of potential outcomes (e.g.,  $\mathbb{E}[Y_0 | X]$ ) could differ between the RCT and observational data, even when the CATE function itself is identified.

#### 3.1 CATE Estimation

The crux of our methodology is to use the CATE estimate from the RCT as a proxy for the true CATE function to falsify or validate an observational estimate. To that end, we first construct an unbiased CATE estimator from RCT data. Since the probability of assignment to each treatment is known by design in RCTs, we can use an IPW-style estimator for the CATE. Similar estimators can be found in standard causal inference textbooks (e.g. Ch. 2 of Hernan and Robins (2021)). A “doubly robust” variant can be used, but if the outcome model is misspecified, this may result in higher variance and a loss of power in our test. Thus, we first define the following “signal” function,

$$\psi_0 = \frac{\mathbf{1}\{S=0\}}{P(S=0|X)} Y \times \left( \frac{\mathbf{1}\{A=1\}}{P(A=1|S=0)} - \frac{\mathbf{1}\{A=0\}}{P(A=0|S=0)} \right), \quad (3)$$

and then observe that the conditional expectation of this signal  $\mathbb{E}[\psi_0 | X]$  is equal to the CATE in the RCT population, using data from the RCT alone.<sup>3</sup>

**Proposition 3.1** (CATE signal from the RCT). *Under Assumption 2.2, the instance-wise CATE signal  $\psi_0$  in Eq. (3), which uses the outcome information from the RCT, is unbiased, i.e.,  $\mathbb{E}[\psi_0|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$ .*

Next, we wish to develop a distinct estimate of the CATE in the RCT population, but one which makes use of the observational data. The first step in building such an estimator is to identify, under our causal assumptions, the corresponding *statistical estimand*, i.e. Eq. (1) in Proposition 2.1. Our goal is to check the *validity* of this estimand, which amounts to challenging Assumptions 2.1 and 2.3. Drawing from existing literature (Dahabreh et al., 2019, 2020; Degtiar and

<sup>3</sup>Note the use of the indicator  $\mathbf{1}\{S=0\}$ , such that  $\psi_0$  only depends on data from the RCT itself, even though we take the conditional expectation over the combined sample.

Rose, 2021), we employ the following doubly robust signal, which combines response surface modeling and inverse probability weighting (IPW):

$$\psi_1 = \frac{1}{P(S=0|X)} \left[ \mathbf{1}\{S=0\} \underbrace{(\mu_1(X) - \mu_0(X))}_{\text{Response Surface Signal}} + \mathbf{1}\{S=1\} \frac{P(S=0|X)}{P(S=1|X)} \underbrace{\left( \frac{\mathbf{1}\{A=1\}(Y - \mu_1(X))}{P(A=1|S=1,X)} - \frac{\mathbf{1}\{A=0\}(Y - \mu_0(X))}{P(A=0|S=1,X)} \right)}_{\text{IPW Signal}} \right], \quad (4)$$

where  $\mu_a(X) := \mathbb{E}[Y | A = a, X, S = 1]$ .

**Proposition 3.2** (CATE signal from the observational data). *Under Assumptions 2.1 and 2.3, the instance-wise CATE signal  $\psi_1$  in Eq. 4, which uses the outcome information from the observational data, is unbiased for the CATE in the RCT population, i.e.,  $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$ .*

We are now ready to give the core result of this section, connecting our causal assumptions to the null hypothesis of the statistical test that we will develop in the next section.

**Corollary 3.1** (Null Hypothesis on Signal Difference). *Define  $\psi = \psi_1 - \psi_0$  as the instance-wise signal difference between the observational and RCT CATE estimates. Then, under the null hypothesis, i.e. under Assumptions 2.1 to 2.3, we have it that  $\mathbb{E}[\psi|X] = 0$ .*

*Proof.* If Assumptions 2.1 to 2.3 hold, then Propositions 3.1 and 3.2 imply that  $\mathbb{E}[\psi_0|X] = \mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$ .  $\square$

**Remark 3.1.** Note that by Corollary 3.1, violation of the conditional moment restrictions imply that one or more of our assumptions is incorrect, including the internal validity assumptions on the RCT. If we are willing to independently assume that the RCT is internally valid, then violation of the CMRs implies a violation of one or both of the internal and external validity assumptions on the observational data.

#### 3.2 Conditional Moment Restriction (CMR) Formulation and Maximum Moment Restriction-based (MMR) Tests

For a practical approach to testing, we leverage the rich literature on conditional moment restriction (CMR) tests. Several examples exist of CMRs being used to express restrictions on functions of the data. One such example is using CMRs to reformulate instrumental variable (IV) regression (Zhang et al., 2020). However, to our knowledge, using CMRs to compare RCT and observational data as described in this paper has not been previously explored. We present the CMR-version of the null hypothesis in the following proposition:

**Proposition 3.3** (Null Hypothesis, CMR). *Under Assumptions 2.1 to 2.3, we have a set of conditional moment restrictions (CMRs) on the signal difference,  $\psi$ :*

$$H_0 : \mathbb{E}[\psi|X] = 0 \quad P_X\text{-almost surely}, \quad (5)$$

where  $P_X$  is the distribution of  $X$  on the joint distribution of the RCT and observational study. Equation (5) implies an infinite set of unconditional moment restrictions,  $\mathbb{E}[\psi f(X)] = 0, \forall f \in \mathcal{F}$ , where  $\mathcal{F}$  is the set of measurable functions on  $\mathcal{X}$ .

The core part of Proposition 3.3 is in showing how we can formulate the CMR given our assumptions, while the second part of the statement is straightforward and follows directly from the law of iterated expectation. Testing CMRs is challenging because an infinite number of equivalent unconditional moment restrictions (UMR) must be considered. Thus, we follow a method proposed by Muandet et al. (2020), where  $\mathcal{F}$  in Proposition 3.3 is set to be a reproducing kernel Hilbert space (RKHS). They further show that using the *maximum moment restriction* (MMR) within the unit ball of the RKHS as the test statistic fully captures the original set of CMRs and also has a closed-form expression that can be easily implemented. However, note that here we are directly testing the CMRs, while Muandet et al. (2020) consider testing hypotheses *on statistical parameters that imply CMRs*, which leads to a larger set of assumptions on the parameters. Therefore, in the following, we state the hypothesis test with respect to the MMR test statistic and the assumptions required for our use case. A proof showing that these assumptions suffice for the properties of the test to hold is provided in Appendix A. This main result will hold for a particular class of kernels, which we define here:

**Definition 3.1** (Integrally strictly positive definite (ISPD)). A kernel  $k(\cdot, \cdot) : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{R}$  is integrally strictly positive definite if for all  $f : \mathcal{W} \rightarrow \mathbb{R}$  satisfying  $0 < \|f\|_2^2 < \infty$ ,

$$\int_{\mathcal{W} \times \mathcal{W}} f(w)k(w, w')f(w')dw dw' > 0$$

Now, we are ready to give an MMR-based hypothesis test that tests the null hypothesis given in Proposition 3.3:

**Theorem 3.1** (Maximum Moment Restriction-based test for CATE function). *Let  $\mathcal{F}$  be a RKHS with reproducing kernel  $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  that is ISPD, continuous and bounded. Suppose  $\|\mathbb{E}[\psi|X]\| < \infty$  almost surely in  $P_X$ , and  $\mathbb{E}[(\psi k(X, X')\psi')^2] < \infty$  where  $(\psi', X')$  is an independent copy of  $(\psi, X)$ . Let  $\mathbb{M}^2 = \sup_{f \in \mathcal{F}, \|f\| \leq 1} (\mathbb{E}[\psi f(X)])^2$ . Then,*

1. *The conditional moment testing problem in Eq. 5 can be reformulated in terms of the MMR as  $H'_0 : \mathbb{M}^2 = 0$ ,  $H'_1 : \mathbb{M}^2 \neq 0$ .*

Further, let the test statistic be the empirical estimate of  $\mathbb{M}^2$ ,

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \psi_i k(x_i, x_j) \psi_j$$

2. Then, under  $H'_0$ ,

$$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

where  $Z_j$  are independent standard normal variables and  $\lambda_j$  are the eigenvalues for  $\psi k(x, x') \psi'$ .

3. Under  $H'_1$ ,

$$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$

where  $\sigma^2 = \text{var}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']]$

*Remark 3.2.* Intuitively, the MMR test statistic is trying to find regions in  $X$  where the signal difference (i.e. the difference in the CATE estimates between the observational study and RCT) is maximized. Thus, the larger the signal difference is, the larger the test statistic will be, and the more likely we will be to reject the null hypothesis. This is exactly the behavior that we want from such a test statistic. Note as well that the test statistic is computed using the signal difference directly and *not* separately for each potential outcome mean. This theoretically-grounded choice follows directly from our discussion in Section 2.2.

*Remark 3.3.* Note that these asymptotic distributions imply that  $n\hat{\mathbb{M}}_n^2$  converges to a distribution with finite variance under the null, but diverges at a rate of  $\sqrt{n}$  under the alternative hypothesis, which implies that the MMR test has asymptotic power of one. In addition, since the null distribution does not have a closed form, to obtain the critical value for rejection, we follow Algorithm 1 in (Muandet et al., 2020), which uses bootstrap to simulate the null distribution.

*Remark 3.4.* Proposition 3.3 states that the true signal difference,  $\psi = \psi_1 - \psi_0$ , satisfies a set of CMRs. However, in practice, we perform testing using an estimate of the signal difference,  $\hat{\psi}$ , where we plug-in estimates of the underlying nuisance functions, such as the propensity score,  $P(A = a|S = 1, X)$ . As a result, we might expect our statistical test, all else being equal, to be more likely to reject an observational study, as there are two sources of variation in the test statistic: first, in the signals themselves through the estimated nuisance functions, and second, through variation in the data that exists even when the signals are perfectly estimated. In our semi-synthetic experiments, we find that the difference in the type I error (between using  $\psi$  and  $\hat{\psi}$ ) is minimal for moderately large sample sizes and converges to zero as the sample size increases.

### 3.3 Explainability of MMR-based Falsification Test

Another appealing feature of using the MMR-based approach is that we may express the maximizer,

$$f^* = \arg \sup_{f \in \mathcal{F}, \|f\| \leq 1} (\mathbb{E}[\psi f(X)])^2, \quad (6)$$

in closed form (see the proof of Corollary 3.2 for details). In turn, we may determine where the CATE function estimated by the RCT and the observational study is the most discrepant by looking at regions of  $\mathcal{X}$  with large magnitudes of  $f^*$ . The function  $f^*$ , known as the “witness function”, can be found by the following corollary:

**Corollary 3.2.** *The witness function in Eq. (6) can be estimated as*

$$\hat{f}^*(x) = C \frac{1}{n} \sum_i \psi_i k(x_i, x)$$

where  $C$  is an unrelated constant so that  $\int_{\mathcal{X}} f^{*2}(x) dx = 1$ .

*Remark 3.5.* Consider the following example where having a witness function could be beneficial: suppose an endocrinologist wants to determine whether to prescribe SGLT2-inhibitors ( $A = 1$ ) or not ( $A = 0$ ) for diabetic patients. Further assume that there is an RCT and an observational study that studies the effect of SGLT2-inhibitors on HbA1c levels ( $Y$ ). If our MMR-based approach were to falsify the observational study, the witness function would enable the clinician to understand what types of patients (e.g. people who are  $\geq 60$  years old and have history of heart disease) have conflicting conclusions in the RCT versus observational study with respect to drug benefit.

With this information, they may seek to understand and do follow-up analyses on what violations of the causal assumptions led to the discrepancy in the “older with prior heart disease” patient population. For example, there may be a violation of internal validity of the observational data (e.g. ignorability), where there are still some unmeasured confounders in the observational study for this particular patient group. Alternatively, there could be an external validity violation (e.g. mean exchangeability of contrast), whereby the causal effects themselves are unbiased, but the standard of care of this patient population may be different between the two studies (i.e. unmeasured effect modifiers). Overall, the witness function can provide a window for clinicians to look for possible violations in a specific patient population, allowing for a richer view into observational study results.

*Remark 3.6.* A practitioner may interpret or visualize the witness function in a couple of different ways, which we outline here. A simple method, appropriate for domain experts (e.g. clinicians), is to use domain knowledge to pre-select a subset of covariates on which one can do low-dimensional projections for each pair. We provide an example of this method in our experimental results. Another method is to take the top or bottom 10% of witness function values over  $\mathcal{X}$  and then look at characteristics of these populations. This

approach can guide the choice of low-dimensional parameters to examine (e.g., for major differences across age, the witness function projected onto age can be plotted).

The MMR-based testing framework is useful both because it affords a closed-form expression of the test statistic used to test the CMR in Proposition 3.3 and gives an explainable view into the rejections of the test via the witness function. We argue that both are crucial for our problem of falsification of causal assumptions in observational studies, specifically internal and external validity. Several other approaches exist in the literature for testing CMRs, and we point the reader to Muandet et al. (2020) for an overview. In the following two sections, we will tease out empirically the benefits of our testing approach against several baselines and provide an example of how the witness function can be used in practice on a real-world dataset.

## 4 SEMI-SYNTHETIC EXPERIMENTS

### 4.1 Setup

For this set of experiments, we use covariates from the Infant Health and Development Program (IHDP), an RCT run on premature infants assessing the treatment effect of professional home visits on future cognitive function (Brooks-Gunn et al., 1992). We generate an RCT and observational dataset (with simulated outcomes) from the partial IHDP dataset used in Hill (2011), which contains 985 observations, 28 covariates, and one binary treatment variable.

A “simulated” dataset in our experiments consists of a single RCT and a single observational dataset. Our simulation strategy for the data draws largely on the approach taken by Hussain et al. (2022). In particular, to generate the RCT, we resample the rows of the IHDP dataset to  $n_0 = 2955$ . For the observational dataset, we first resample the rows of the IHDP dataset to the desired sample size,  $n = s \cdot n_0$ . Then, we induce a difference in the covariate distribution between the observational component and the RCT by doing weighted resampling in the observational data, such that male infants, infants whose mothers smoked, and infants with working mothers are less prevalent. To introduce explicit violations of our assumptions in the observational data, we generate  $m$  confounders so that we can later conceal some of them to simulate unmeasured confounding. Then, in both the RCT and the observational dataset, we simulate outcomes according to a response surface detailed in Appendix C. Finally, we conceal  $c_z$  confounders in order of “confounding strength”, which is determined by a vector,  $\gamma \in \mathbb{R}^m$ . For more information on confounder generation, outcome simulation, and bias simulation via confounder concealment, see Appendix C. For parameters  $m$ ,  $c_z$ , and  $\alpha$  (significance level), we default to  $m = 7$ ,  $c_z = 0$ , and  $\alpha = 0.05$  unless otherwise specified.

## 4.2 Evaluation

We evaluate our algorithm based on our original desiderata. Namely, we measure *power*, i.e. the rate of rejecting the null hypothesis when the CATE function estimates from the RCT and observational study *do not* converge to the same function and *type I error*: rate of rejecting the null hypothesis given that they *do* converge to the same function.

We use the following two baselines. **Average Treatment Effect (ATE)** – in the RCT, we compute the difference of mean outcomes between the treatment and control groups; in the observational data, we obtain an ATE estimate by leveraging recent techniques in the double machine learning (DML) and transportability literature, akin to the estimator in Dahabreh et al. (2020). **Group Average Treatment Effect (GATE)** – in the RCT, we compute the difference of mean outcomes between the treatment and control groups in pre-specified subgroups defined by the infant’s birth weight and maternal marital status<sup>4</sup>; in the observational data, we use a transportable, doubly-robust estimator (see Appendix C in Hussain et al. (2022)), to estimate the GATE for each subgroup. Both baselines use hypothesis testing based on asymptotic normality of ATE or subgroup estimates. Note that this approach requires pre-specification of the subgroups.

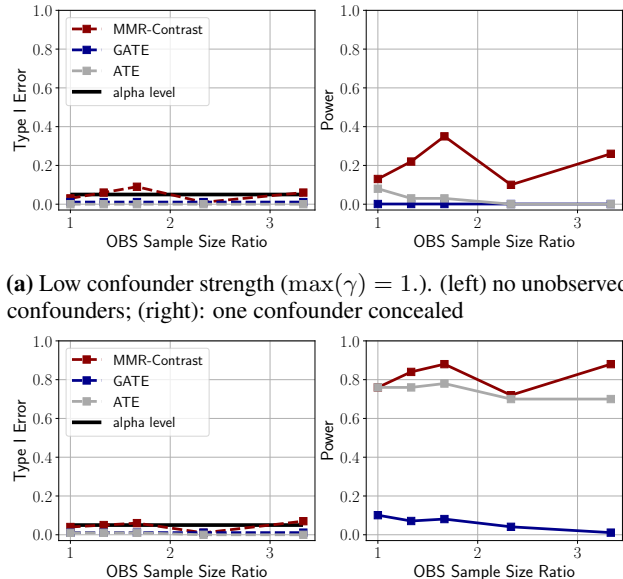
Both baselines reflect the idea of “falsifying” the observational study by looking at a pre-specified group (subgroups, in the GATE case) to detect differences in the causal effect estimates. Our method, labeled as **MMR-Contrast**, requires no pre-specification and automatically finds “highly-discrepant” regions where the causal effect estimate is different between the RCT and the observational study.

## 4.3 Results

*MMR-Contrast largely maintains the desired type I error of 0.05 while having more power compared to baselines. As conjectured in Remark 3.4, MMR-Contrast tends to slightly over-reject, which is reflected in Fig. 1 by the marginally elevated type I error. Furthermore, MMR-Contrast enjoys greater power than GATE and ATE, particularly in settings where the confounding bias is more subtle. We conjecture that the gain in power is due to MMR-Contrast implicitly checking across all subpopulations of  $X$  for disagreements in CATE estimates. Indeed, we see that when the concealed confounder has a weight of 1 (as opposed to 2.75), the difference in power between MMR and ATE is much larger.*

*When computing MMR-Contrast with  $\psi$  versus  $\hat{\psi}$ , the empirical gap in type I error shrinks with increasing sample size of the observational study (see Fig. 2a). Reassuringly, we see that the level of the test is maintained at  $\alpha = 0.05$  when the true signal difference,  $\psi$ , is used, which supports our theoretical results. Secondly, using the estimated signal*

<sup>4</sup>We specify the following four subgroups: ( $\geq 2000g$ , married), ( $< 2000g$ , married), ( $\geq 2000g$ , single), ( $< 2000g$ , single)



(a) Low confounder strength ( $\max(\gamma) = 1.$ ). (left) no unobserved confounders; (right): one confounder concealed

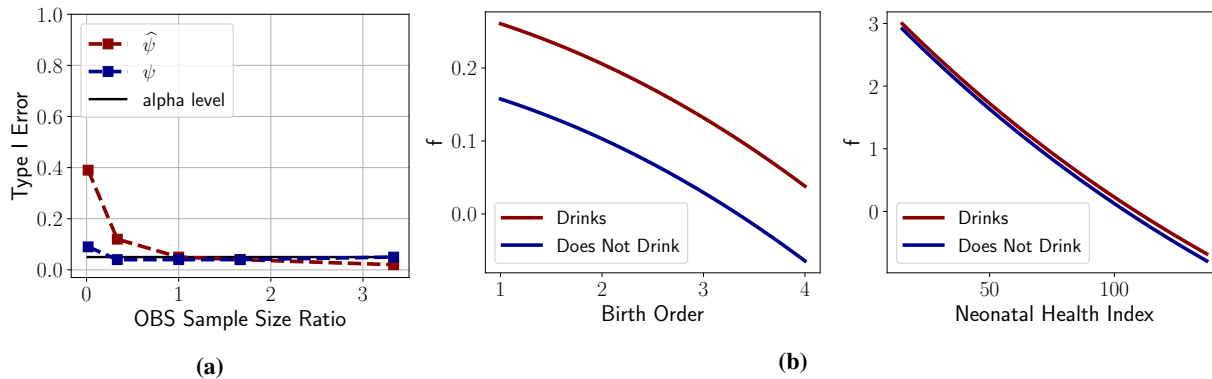
(b) High confounder strength ( $\max(\gamma) = 2.75$ ). (left) no unobserved confounders; (right): one confounder concealed

**Figure 1:** Type I error and power of MMR-Contrast, GATE and ATE under different confounder strengths. The left panels in a) and (b) show that the level of all three approaches generally retains the nominal level of 0.05. The right panels show the superior power of MMR-Contrast. Particularly, when the confounder strength is lower (as in (a)), the difference of CATE estimates between the observational study and the RCT is more difficult to detect, leading to a larger difference of power between MMR-Contrast and ATE. The GATE approach, since it is based on random subgroups, has minimal power, even under the high confounder strength scenario.

difference,  $\hat{\psi}$ , achieves the appropriate type I error when the observational study size at least matches the RCT, i.e. sample size ratio is 1, which one might expect in practice.

*Visualizing the witness function in Fig. 2b demonstrates the covariate regions in which the observational effect estimates are increased or decreased compared to the RCT. We largely see that the witness function yields positive values, implying that the observational study is generally estimating a larger treatment effect (i.e. professional visits benefit child cognitive development) than the RCT. However, there are certain subgroups, e.g. children with high birth order whose mothers do not drink and children with high neonatal health index, for which the observational study estimates lower treatment effects than the RCT. The MMR test is able to discover these subgroup differences, leading to better power than testing for ATE or GATE. Another potential use case of the witness function is for development of treatment guidelines, where subgroups with high witness function values may be “flagged” as having conflicting evidence.*





**Figure 2:** Panel (a) demonstrates the relative performance of tests using test statistics computed with true signals ( $\psi$ ) and estimated signals ( $\hat{\psi}$ ). The sample size of the RCT study is fixed ( $n_0 = 2955$ ) and the sample size ratio between the observational study and the original IHDP data ranges from 0.01 to 3.33. The blue line shows that the test using  $\psi$  achieves the nominal level (0.05). The red line shows that under small sample sizes of the observational study, the test using  $\hat{\psi}$  over-rejects due to errors in nuisance function estimation, which is consistent with our conjecture. Nevertheless, its level promptly converges to 0.05 as the number of samples in the observational study matches or exceeds the RCT. Panel (b) demonstrates the witness functions produced as a byproduct of our test, which show mostly positive values and certain negative regions.

Selection Bias	MMR-Contrast	ATE	GATE
$p = 0$	0.29	0.32	<b>0.17</b>
$p = 0.05$	<b>0.67</b>	0.58	0.40
$p = 0.10$	<b>0.94</b>	0.88	0.67
$p = 0.15$	<b>1.0</b>	0.98	0.91

**Table 1:** Rejection rate when introducing different amounts of selection bias into the observational data in WHI study.  $p$  stands for the strength of selection introduced in the data (refer to Section 5 for details).

## 5 WOMEN’S HEALTH INITIATIVE (WHI) EXPERIMENTS

To assess our method in a practical setting, we use observational and clinical trial data from the Women’s Health Initiative (WHI). These studies broadly investigate the impact of hormone therapy and vitamin D supplementation on several clinical outcomes. Conceptually, our analysis consists of first taking  $B$  bootstrapped datasets from either the original WHI observational study or a “biased” version, then selecting a subset of covariates (to generate subgroups for the GATE baseline), and finally running *GATE*, *ATE*, and *MMR-Contrast* on each bootstrap iteration. We evaluate the methods by reporting the *rejection rate* in each “bias” setting. To induce different amounts of selection bias into the observational study, we drop patients who were not exposed to the intervention and did not experience the event with some probability  $p$ . For further details on data preprocessing, setup, and evaluation, see Appendix D.

### 5.1 Setup

We use the Postmenopausal Hormone Therapy (PHT) trial as the RCT in our analysis, which was run on postmenopausal women aged 50-79 years with an intact uterus. The trial

investigated the effect of hormone therapy on several types of cancers, cardiovascular events, and fractures, measuring the “time-to-event” for each outcome. In the WHI setup, the observational study component was run in parallel and tracked similar outcomes to the RCT. Our processing of this dataset follows closely to the pre-processing steps taken by Hussain et al. (2022). We binarize a composite outcome, called the “global index”, in our analysis, where  $Y = 1$  if coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes was observed in the first seven years of follow-up, and  $Y = 0$  otherwise. Note that  $Y = 0$  could also occur from censoring. To establish treatment and control groups in the observational study, we use questionnaire data in which participants confirm or deny usage of combination hormones (i.e. both estrogen and progesterone) in the first three years. For other covariates, we use only those measured in both the RCT and observational study to simplify the analysis.

### 5.2 Results

*MMR-Contrast has superior power compared to the baselines in real-world data.* As shown in Table 1, MMR-Contrast has the best ability to reject studies that have selection bias. Note, as well, that GATE always has lower rejection probability compared to ATE. This result implies that using the GATE approach without prior knowledge on which subgroups lead to different effect estimates using observational versus RCT data is highly disadvantageous in terms of statistical power. Though the MMR approach is conceptually similar to GATE, it finds discrepant covariate regions in a data-driven fashion instead of requiring pre-specified groups, thus achieving better power.



## 6 RELATED WORK

**Transportability of causal effects:** A long line of work gives assumptions under which causal effect estimates can be transported from one population to another. This includes work in statistics on generalizing average effects from one or more RCTs to broader target populations (Cole and Stuart, 2010; Hartman et al., 2015; Dahabreh et al., 2019, 2020), and work in computer science on giving graphical criteria for determining when effects can be transported in more general scenarios (Pearl and Bareinboim, 2011, 2014; Pearl, 2015). Hartman et al. (2015) similarly consider hypothesis testing as a “placebo” test to check assumptions, though their assumptions differ from the ones we consider here. They focus on transporting effects from RCTs to a target population, and do not assume internal validity on observational data. In particular, their assumptions have a testable implication, that the average treated outcome in the target population will match the average treated outcome under a reweighting of the RCT population. Meanwhile, our focus is on testing assumptions of both transportability / external validity, as well as internal validity of observational studies.

**Combining experimental and observational data for improved estimation:** There is a recent line of work on combining observational and experimental data to yield more precise estimates of causal effects, even when the observational data may be biased (Rosenman et al., 2020; Yang et al., 2020; Cheng and Cai, 2021; Chen et al., 2021). We focus on hypothesis testing as a means of falsification as our primary goal rather than merging data. While Yang et al. (2020) use hypothesis testing as a part of their approach, their test depends on the parametric form of the CATE function that they seek to estimate, while our test is nonparametric in nature.

**Minimax and variational methods for parameter estimation via CMRs:** In addition, there is a growing literature of minimax algorithms that aim to find a well-specified set of model parameters that fulfill a set of CMRs. For example, one line of work looks at constructing a minimax formulation of the generalized method of moments (GMM) framework that aims to estimate highly non-linear model parameters that are also solutions to the moment conditions implied by the problem at hand, e.g. IV regression (Lewis and Syrgkanis, 2018; Dikkala et al., 2020; Bennett et al., 2019; Bennett and Kallus, 2020). Metzger (2022) also gives asymptotic theory for minimax estimators of functionals in CMRs. Another line of work tackles the case where a set of CMRs may only weakly identify the nuisance functions, though the target parameter may still be efficiently and uniquely estimable under some conditions on the estimator (Kallus et al., 2021; Bennett et al., 2022). This literature focuses on the problem of parameter estimation, while we use the CMR formulation for hypothesis testing of transportability and internal validity assumptions.

## 7 DISCUSSION & LIMITATIONS

We have proposed a novel approach for falsifying the assumptions of observational studies using experimental data. These causal assumptions, stating the internal and external validity of observational and experimental data, imply that the conditional average treatment effect is equivalent across all observed subpopulations of  $X$  in both the observational and experimental data. This in turn gives rise to testable restrictions on the combined data distribution, implying, as we show, that the difference between two functions of the data is zero-mean for any subset of  $X$ . Recent advances in the econometrics literature allow us to test such restrictions. Our approach implicitly searches for regions of  $X$  where the CATE estimates disagree between the observational and experimental data, without the need for pre-specifying these subpopulations. Moreover, this approach yields a function that characterizes the regions where disagreement is large. Finally, we design our test to avoid rejecting studies due to differences in baseline factors that do not influence the treatment effect.

However, our approach shares certain limitations with some methods in the literature on testing for violation of causal assumptions. In particular, while violations of the CMRs imply violations of causal assumptions, this does not directly tell us which assumptions are violated (e.g., whether the observational study is subject to hidden confounding, or whether there is simply an unobserved difference between the RCT and observational populations). Finally, due to the fact that policy guidelines can be formed from such RCTs and observational studies in society, it is important for practitioners to consider the biases in the data as well as the aforementioned limitation of our method.

### Acknowledgments

We would like to thank members of the Clinical Machine Learning group for helpful discussions and valuable feedback on the manuscript. ZH was supported by an ASPIRE award from The Mark Foundation for Cancer Research and by the National Cancer Institute of the National Institutes of Health under Award Number F30CA268631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MCS was supported by the LEAP program from the Ministry of Science and Technology in Taiwan. MO and DS were supported in part by Office of Naval Research Award No. N00014-21-1-2807. ID is supported by an Eric and Wendy Schmidt Center PhD fellowship.

### References

Usama Ahmed Ali, R Joren, Beata M Reiber, Pieter C van der Sluis, and Marc G Besselink. Sample size of surgical randomized controlled trials: a lack of improvement over time. *Journal of Surgical Research*, 228:1–7, 2018.

- Lindsey R Baden, Hana M El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, Stephen A Spector, Nadine Rouphael, C Buddy Creech, et al. Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England journal of medicine*, 2020.
- Andrew Bennett and Nathan Kallus. The variational method of moments. *arXiv preprint arXiv:2012.09422*, 2020.
- Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.
- Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *arXiv e-prints*, pages arXiv–2208, 2022.
- Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.
- Shuxiao Chen, Bo Zhang, and Ting Ye. Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint (2109.10522)*, September 2021.
- David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint (2111.15012)*, November 2021.
- Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American journal of epidemiology*, 172(1):107–115, July 2010.
- Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A Katz, Miguel A Hernán, Marc Lipsitch, Ben Reis, and Ran D Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 2021.
- Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019.
- Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.
- Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*, 2021.
- Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.
- Jessica M Franklin, Elisabetta Patorno, Rishi J Desai, Robert J Glynn, David Martin, Kenneth Quinto, Ajinkya Pawar, Lily G Bessette, Hemin Lee, Elizabeth M Garry, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the rct duplicate initiative. *Circulation*, 143(10):1002–1013, 2021.
- Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A*, 178(3):757–778, June 2015.
- Miguel A Hernan and James M Robins. *Causal Inference*. CRC Press, Boca Raton, FL, February 2021.
- Miguel A Hernán, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Meir J Stampfer, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6):766, 2008.
- Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.
- Zeshan Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. *arxiv preprint arXiv:2209.13708*, 2022.
- Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.
- Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.
- Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.
- Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.
- Jonas Metzger. Adversarial estimators. *arXiv preprint arXiv:2204.10495*, 2022.
- Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.
- Judea Pearl. Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266, September 2015.
- Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254, August 2011.

- Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, November 2014.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Rachel Phillips, Lorna Hazell, Odile Sauzet, and Victoria Cornelius. Analysis and reporting of adverse events in randomised controlled trials: a review. *BMJ open*, 9(2): e024537, 2019.
- Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*, 2020.
- Jacques E Rossouw, Garnet L Anderson, Ross L Prentice, Andrea Z LaCroix, Charles Kooperberg, Marcia L Stefanick, Rebecca D Jackson, Shirley AA Beresford, Barbara V Howard, Karen C Johnson, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women’s health initiative randomized controlled trial. *Jama*, 288(3):321–333, 2002.
- Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.
- Robert J Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, September 2009.
- SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015.
- Ruth Tsang, Lindsey Colley, and Larry D Lynd. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of clinical epidemiology*, 62(6):609–616, 2009.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.
- Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY, 2004.
- Yu Xie, Christopher Near, Hongwei Xu, and Xi Song. Heterogeneous treatment effects on children’s cognitive/non-cognitive skills: A reevaluation of an influential early childhood intervention. *Social science research*, 86: 102389, 2020.
- Shu Yang, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomized trial and Real-World data for treatment heterogeneity estimation. *arXiv preprint (2005.10579)*, May 2020.
- Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020.

## APPENDIX

### A Proofs

#### A.1 Proof of Proposition 2.1

**Proposition 2.1.** *Under Assumptions 2.1 and 2.3, the CATE of the RCT given  $X$ ,  $\mathbb{E}[Y_1 - Y_0 | X, S = 0]$ , is identifiable in the observational data by*

$$\mathbb{E}[Y | X, A = 1, S = 1] - \mathbb{E}[Y | X, A = 0, S = 1] \quad (1)$$

*Proof.*

$$\begin{aligned} & \mathbb{E}[Y_1 - Y_0 | X, S = 0] \\ &= \mathbb{E}[Y_1 - Y_0 | X, S = 1] \\ &= \mathbb{E}[Y_1 | X, S = 1] - \mathbb{E}[Y_0 | X, S = 1] \\ &= \mathbb{E}[Y | X, A = 1, S = 1] - \mathbb{E}[Y | X, A = 0, S = 1] \end{aligned}$$

The first equality follows from the mean exchangeability of the contrast (Assumption 2.3) and the second from the linearity of the expectation operator. The final equality follows from ignorability and consistency (Assumption 2.1).  $\square$

#### A.2 Proof of Proposition 3.1

**Proposition 3.1** (CATE signal from the RCT). *Under Assumption 2.2, the instance-wise CATE signal  $\psi_0$  in Eq. (3), which uses the outcome information from the RCT, is unbiased, i.e.,  $\mathbb{E}[\psi_0 | X] = \mathbb{E}[Y_1 - Y_0 | X, S = 0]$ .*

*Proof.* Here, we show that the signal

$$\psi_0 = \left( \frac{\mathbf{1}\{A = 1\}}{P(A = 1 | S = 0)} - \frac{\mathbf{1}\{A = 0\}}{P(A = 0 | S = 0)} \right) \cdot \frac{\mathbf{1}\{S = 0\}}{P(S = 0 | X)} Y$$

is an unbiased estimator of the CATE in RCT population under consistency and fully randomized treatment assignment (i.e.,  $P(A | X, S = 0) = P(A | S = 0)$ , and  $Y_a \perp\!\!\!\perp A$  as in Assumption 2.2). In particular, we can observe that

$$\begin{aligned} \mathbb{E}[\psi_0(A, Y, S, X) | X] &= \sum_{A, Y, S} \psi_0(A, Y, S, X) \cdot P(A, Y, S | X) \\ &= \sum_{A, Y, S} \frac{\mathbf{1}\{A = 1\}}{P(A = 1 | S = 0)} \cdot \frac{\mathbf{1}\{S = 0\}}{P(S = 0 | X)} Y \cdot P(A, Y, S | X) \\ &\quad - \sum_{A, Y, S} \frac{\mathbf{1}\{A = 0\}}{P(A = 0 | S = 0)} \cdot \frac{\mathbf{1}\{S = 0\}}{P(S = 0 | X)} Y \cdot P(A, Y, S | X) \end{aligned} \quad (7)$$

We focus on the first term, observing that the second term can be handled similarly. We first re-write the first term as

$$\begin{aligned} & \sum_{A, Y, S} \frac{P(Y | S, A, X) P(A | S, X) P(S | X)}{P(A = 1 | S = 0) P(S = 0 | X)} \cdot \mathbf{1}\{A = 1, S = 0\} \cdot Y \\ &= \sum_Y Y \cdot \frac{P(Y | S = 0, A = 1, X) P(A = 1 | S = 0, X) P(S = 0 | X)}{P(A = 1 | S = 0) P(S = 0 | X)} \\ &= \mathbb{E}[Y | S = 0, A = 1, X] \\ &= \mathbb{E}[Y_1 | S = 0, A = 1, X] \\ &= \mathbb{E}[Y_1 | X, S = 0] \end{aligned} \quad (8)$$

Repeating the similar arguments for the second term in Eq. 7, we have  $\mathbb{E}[\psi_0(A, Y, S, X) | X] = \mathbb{E}[Y_1 - Y_0 | X, S = 0]$ , which completes the proof.  $\square$

### A.3 Proof of Proposition 3.2

**Proposition 3.2** (CATE signal from the observational data). *Under Assumptions 2.1 and 2.3, the instance-wise CATE signal  $\psi_1$  in Eq. 4, which uses the outcome information from the observational data, is unbiased for the CATE in the RCT population, i.e.,  $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S = 0]$ .*

*Proof.* We have,

$$\psi_1 = \frac{1}{P(S=0|X)} \left[ \mathbf{1}\{S=0\} (\mu_1(X) - \mu_0(X)) + \mathbf{1}\{S=1\} \frac{P(S=0|X)}{P(S=1|X)} \left( \frac{\mathbf{1}\{A=1\} (Y - \mu_1(X))}{P(A=1|S=1,X)} - \frac{\mathbf{1}\{A=0\} (Y - \mu_0(X))}{P(A=0|S=1,X)} \right) \right]$$

$$\begin{aligned} \mathbb{E}[\psi_1(A, Y, S, X) | X] &= \frac{1}{P(S=0|X)} \left[ \sum_{A,Y} (\mu_1(X) - \mu_0(X)) P(Y | S=0, A, X) P(A | S=0, X) P(S=0|X) \right. \\ &+ \frac{P(S=0|X)}{P(S=1|X)} \left( \sum_Y \frac{Y - \mu_1(X)}{P(A=1|S=1,X)} P(Y | S=1, A=1, X) P(A=1|S=1,X) P(S=1|X) \right. \\ &\left. \left. - \sum_Y \frac{Y - \mu_0(X)}{P(A=0|S=1,X)} P(Y | S=1, A=0, X) P(A=0|S=1,X) P(S=1|X) \right) \right] \\ &= \sum_{A,Y} (\mu_1(X) - \mu_0(X)) P(Y | S=0, A, X) P(A | S=0, X) \\ &\quad + \sum_Y (Y - \mu_1(X)) P(Y | S=1, A=1, X) \\ &\quad - \sum_Y (Y - \mu_0(X)) P(Y | S=1, A=0, X) \\ &= (\mu_1(X) - \mu_0(X)) \underbrace{\sum_{A,Y} P(Y, A | S=0, X)}_{=1} \\ &\quad + \mathbb{E}[Y | S=1, A=1, X] - \mu_1(X) \cdot \underbrace{\sum_Y P(Y | S=1, A=1, X)}_{=1} \\ &\quad - \mathbb{E}[Y | S=1, A=0, X] - \mu_0(X) \cdot \underbrace{\sum_Y P(Y | S=1, A=0, X)}_{=1} \tag{9} \\ &= \mathbb{E}[Y | S=1, A=1, X] - \mathbb{E}[Y | S=1, A=0, X] \tag{10} \\ &= \mathbb{E}[Y_1 - Y_0 | X, S=0] \tag{11} \end{aligned}$$

Note that we have Eq. 9 and Eq. 10 since  $\mu_a(X) = \mathbb{E}[Y | S=1, A=a, X]$ . Eq. 11 follows from Proposition 2.1.  $\square$

### A.4 Proof of Proposition 3.3

We restate Proposition 3.3 here for convenience.

**Proposition 3.3** (Null Hypothesis, CMR). *Under Assumptions 2.1 to 2.3, we have a set of conditional moment restrictions (CMRs) on the signal difference,  $\psi$ :*

$$H_0 : \mathbb{E}[\psi|X] = 0 \quad P_X\text{-almost surely,} \tag{5}$$

where  $P_X$  is the distribution of  $X$  on the joint distribution of the RCT and observational study. Equation (5) implies an infinite set of unconditional moment restrictions,  $\mathbb{E}[\psi f(X)] = 0, \forall f \in \mathcal{F}$ , where  $\mathcal{F}$  is the set of measurable functions on  $\mathcal{X}$ .

*Proof.* Under Assumptions 2.1 to 2.3, we have  $\mathbb{E}[\psi_0|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$  and  $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$  by Propositions 3.1 and 3.2 as discussed in Section 3.1. That is,  $\mathbb{E}[\psi|X] = 0$  where  $\psi = \psi_1 = \psi_0$  is the signal difference given

two CATE signals. Let  $\mathcal{F}$  be the set of measurable functions on  $\mathcal{X}$ . Then, by the Law of Iterated Expectations we have

$$\mathbb{E}[\psi f(X)] = \mathbb{E}_X[\mathbb{E}[\psi f(X)|X]] = \mathbb{E}_X[\mathbb{E}[\psi|X]f(X)], \quad P_X\text{-a.s.}, \forall f \in \mathcal{F}$$

We see that Eq. 5 implies the following infinite set of unconditional moment restrictions,

$$\mathbb{E}[\psi f(X)] = 0, \quad P_X\text{-a.s.}, \forall f \in \mathcal{F}$$

□

### A.5 Proofs for Theorem 3.1 and Corollary 3.2

We restate the theorem and corollary here for convenience.

**Theorem 3.1** (*Maximum Moment Restriction-based test for CATE function*). *Let  $\mathcal{F}$  be a RKHS with reproducing kernel  $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  that is ISPD, continuous and bounded. Suppose  $|\mathbb{E}[\psi|X]| < \infty$  almost surely in  $P_X$ , and  $\mathbb{E}[[\psi k(X, X')\psi']^2] < \infty$  where  $(\psi', X')$  is an independent copy of  $(\psi, X)$ . Let  $\mathbb{M}^2 = \sup_{f \in \mathcal{F}, \|f\| \leq 1} (\mathbb{E}[\psi f(X)])^2$ . Then,*

1. *The conditional moment testing problem in Eq. 5 can be reformulated in terms of the MMR as  $H'_0 : \mathbb{M}^2 = 0$ ,  $H'_1 : \mathbb{M}^2 \neq 0$ .*

Further, let the test statistic be the empirical estimate of  $\mathbb{M}^2$ ,

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \psi_i k(x_i, x_j) \psi_j$$

2. *Then, under  $H'_0$ ,*

$$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

where  $Z_j$  are independent standard normal variables and  $\lambda_j$  are the eigenvalues for  $\psi k(x, x')\psi'$ .

3. *Under  $H'_1$ ,*

$$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$

where  $\sigma^2 = \text{var}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']]$

**Corollary 3.2.** *The witness function in Eq. (6) can be estimated as*

$$\hat{f}^*(x) = C \frac{1}{n} \sum_i \psi_i k(x_i, x)$$

where  $C$  is an unrelated constant so that  $\int_{\mathcal{X}} f^{*2}(x) dx = 1$ .

The following proof follows (Muandet et al., 2020). Let us define the following operator,

$$Mf = \mathbb{E}[\psi f(X)] \tag{12}$$

where  $f \in \mathcal{F}$ . Since  $|\mathbb{E}[\psi|X]| < \infty$  almost surely in  $P_X$ ,  $M$  is a bounded linear operator. By Riesz representation theorem, there exists a unique  $g \in \mathcal{F}$  such that

$$Mf = \langle f, g \rangle$$

where

$$g = \mathbb{E}[\psi k(X, \cdot)].$$

$g$  is called the conditional moment embedding (CMME) of the CMR,  $\mathbb{E}[\psi|X]$ , in  $\mathcal{F}$  w.r.t.  $P_X$ . Therefore, it follows that

$$\mathbb{M}^2 = \sup_{f \in \mathcal{F}, \|f\| \leq 1} (\mathbb{E}[\psi f(X)])^2 = \sup_{f \in \mathcal{F}, \|f\| \leq 1} \langle f, g \rangle^2 = \left\langle \frac{g}{\|g\|}, g \right\rangle^2 = \|g\|^2$$

Note that the above implies that the witness function  $f^* = \arg \sup_{f \in \mathcal{F}, \|f\| \leq 1} (\mathbb{E}[\psi f(X)])^2 = \frac{g}{\|g\|}$ . Since  $g$  is defined as  $\mathbb{E}[\psi k(X, \cdot)]$ , it can be empirically estimated as  $\frac{1}{n} \sum_{i=1}^n \psi_i k(x_i, \cdot)$ , which leads to Corollary 3.2.

Since  $\mathbb{M}^2 = \|g\|^2$ , the first statement in Theorem 3.1 is essentially

$$\mathbb{E}[\psi|X] = 0, P_X\text{-almost surely} \Leftrightarrow \|g\|^2 = 0$$

That is,  $g \in \mathcal{F}$  fully captures the information of the CMR for all  $x \in \mathcal{X}$ . This equivalence, which we will now prove, is crucial since our statistical test is based on  $\|g\|^2$  and its estimates, while Proposition 3.3 is directed to the CMR:

( $\Rightarrow$ ) We note that since  $\mathcal{F}$  is a Hilbert space, it follows that  $g \in \mathcal{F}$ , and  $\forall f \in \mathcal{F}, \langle f, g \rangle = \mathbb{E}[\psi f(X)] = 0$  (Proposition 3.3).  $g$  can now only be a zero vector. Therefore,  $\|g\|^2 = 0$ .

( $\Leftarrow$ )

$$\begin{aligned} & \|g\|^2 = 0 \\ \Rightarrow & \|\mathbb{E}[\psi k(X, \cdot)]\|^2 = 0 \\ \Rightarrow & \|\mathbb{E}[\mathbb{E}[\psi|X]k(X, \cdot)]\|^2 = 0 \\ \Rightarrow & \left\| \int_{\mathcal{X}} k(x, \cdot) \mathbb{E}[\psi|x] p_X(x) dx \right\|^2 = 0 \\ \Rightarrow & \iint_{\mathcal{X} \times \mathcal{X}} p_X(x) \mathbb{E}[\psi|x] k(x, x') \mathbb{E}[\psi|x'] p_X(x') dx dx' = 0 \\ \Rightarrow & \|\mathbb{E}[\psi|x] p_X(x)\|^2 = 0 \quad (\because k(\cdot, \cdot) \text{ is ISPD}) \\ \Rightarrow & \mathbb{E}[\psi|x] = 0, P_X\text{-almost surely} \end{aligned}$$

Finally, we move to the second and third statements of Theorem 3.1, which define the estimator and its statistical properties. Since  $\mathbb{M}^2 = \|g\|^2 = \|\mathbb{E}[\psi k(X, \cdot)]\|^2 = \mathbb{E}[\mathbb{E}[\psi k(X, X')\psi']]$  where  $(X, \psi)$  and  $(X', \psi')$  are independently and identically distributed, we may use a  $U$ -statistic to estimate  $\mathbb{M}^2$ , which is exactly

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \psi_i k(x_i, x_j) \psi_j$$

The asymptotic distribution of  $U$ -statistics has been investigated intensively in the literature. Specifically, from Section 5.5 of Serfling (2009), taking the special case of kernels with two inputs, we have the following lemma:

**Lemma A.1** ((Serfling)). *Given a kernel  $h(\cdot, \cdot) : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{R}$  where  $\mathbb{E}_{(W, W')} [h(W, W')] = \theta$  and  $\mathbb{E}_{(W, W')} [h^2(W, W')] < \infty$ , the asymptotic distribution of the  $U$ -statistic  $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} h(w_i, w_j)$  can be categorized into two cases based on  $\zeta_1 = \text{var}_W(\mathbb{E}_{W'} [h(W, W')])$ :*

$$\begin{cases} \sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, 4\zeta_1), & \zeta_1 > 0 \\ n(U_n - \theta) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1), & \zeta_1 = 0 \end{cases}$$

where  $Z_j$  are independent standard normal variables and  $\lambda_j$  are the eigenvalues of  $h$ , i.e. the solutions for  $\mathbb{E}_{W'} [h(W', w)v(w)] - \lambda v(w) = 0$

Note that if we set  $W = (\psi, X)$ ,  $h(W, W') = \psi k(X, X')\psi'$ ,  $\theta = \mathbb{M}^2$ ,  $\zeta_1 = \sigma^2$ , the second and third statements of Theorem 3.1 holds as long as  $\mathbb{M}^2 = 0 \Leftrightarrow \sigma^2 = \text{var}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']] = 0$ , which we will now show:

( $\Rightarrow$ )

$$\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi'] = \langle \psi k(X, \cdot), \mathbb{E}_{(\psi', X')}[\psi' k(X', \cdot)] \rangle = \|\psi k(X, \cdot)\| \left\langle \frac{\psi k(X, \cdot)}{\|\psi k(X, \cdot)\|}, g \right\rangle$$

Now since

$$\frac{\psi k(X, \cdot)}{\|\psi k(X, \cdot)\|} \in \mathcal{F}, \left\| \frac{\psi k(X, \cdot)}{\|\psi k(X, \cdot)\|} \right\| = 1$$

and

$$\mathbb{M}^2 = 0 \Rightarrow \sup_{f \in \mathcal{F}, \|f\| \leq 1} \langle f, g \rangle = 0 \Rightarrow \langle f, g \rangle = 0, \forall f \in \mathcal{F}, \|f\| \leq 1$$

We conclude  $\mathbb{M}^2 = 0 \Rightarrow \left\langle \frac{\psi k(X, \cdot)}{\|\psi k(X, \cdot)\|}, g \right\rangle = 0 \Rightarrow \mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi'] = 0 \Rightarrow \text{var}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']] = 0$



( $\Leftarrow$ )

We first note that  $\text{var}_{(\psi, X)}(\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']) = 0$  implies that  $\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']$  is a constant  $P_{(\psi, X)}$ -almost surely. We denote this constant as  $c$  so we have

$$\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi'] = c, P_{(\psi, X)}\text{-almost surely} \quad (13)$$

From the definition of  $\psi$ , let  $X = x^*$  be in the support of the observational study, then

$$\begin{aligned} \mathbb{E}[\psi|S = 1, X = x^*] &= \frac{1}{P(S = 1|X = x^*)} \mathbb{E} \left[ \frac{\mathbf{1}(A = 1)(Y - \mu_1(x^*))}{P(A = 1|S = 1, X = x^*)} \right. \\ &\quad \left. - \frac{\mathbf{1}(A = 0)(Y - \mu_0(x^*))}{P(A = 0|S = 1, X = x^*)} | S = 1, X = x^* \right] \\ &= \frac{1}{P(S = 1|X = x^*)} \left[ \mathbb{E}[Y - \mu_1(x^*)|A = 1, S = 1, X = x^*] \right. \\ &\quad \left. - \mathbb{E}[Y - \mu_0(x^*)|A = 0, S = 1, X = x^*] \right] \\ &= 0, \end{aligned}$$

where the last equality stems from the definition of  $\mu_1$  and  $\mu_0$ . Now note that

$$\begin{aligned} \mathbb{E}_{\psi}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi'|S = 1, X = x^*]] &= \mathbb{E}_{\psi}[\mathbb{E}_{(\psi', X')}[\psi k(x^*, X')\psi'|S = 1, X = x^*]] \\ &= \mathbb{E}_{\psi}[\psi|S = 1, X = x^*]\mathbb{E}_{(\psi', X')}[k(x^*, X')\psi'] \\ &= 0 \cdot \mathbb{E}_{(\psi', X')}[k(x^*, X')\psi'] = 0 \end{aligned}$$

But also we have, from (13),

$$\mathbb{E}_{\psi}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi'|S = 1, X = x^*]] = \mathbb{E}_{\psi}[c] = c$$

Therefore, we have  $c = 0$  and thus

$$\mathbb{M}^2 = \mathbb{E}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi k(X, X')\psi']] = \mathbb{E}_{(\psi, X)}[0] = 0$$

so this side of the arrow is also proven.

## B Motivating Empirical Examples of Generalization of Treatment Effects rather than Counterfactual Means

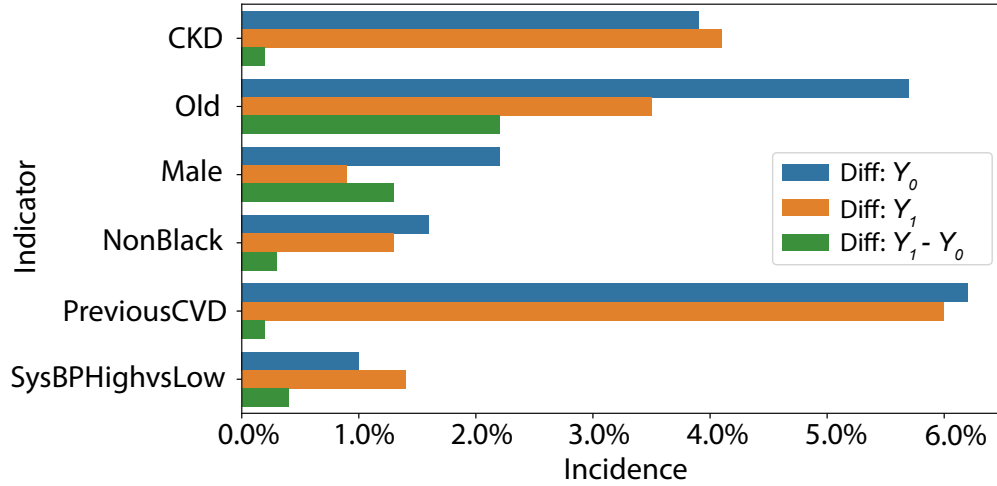
In Fig. 3 we plot data that is publicly available in SPRINT Research Group (2015) and Franklin et al. (2021). The former is a randomized trial that reports on outcomes across subgroups, where we observe that subgroups often have larger differences in their baseline outcomes than in their treatment effects. The latter is a study that attempts to replicate ten RCTs using observational data. For each observational study and trial, they report on not only the resulting differences in rates (between treatment and control), but also the marginal rates under each of treatment and control. This is done for both the observational studies and the original RCTs. We can observe that the estimated ‘‘treatment effects’’ tend to be closer together (between the observational studies and RCTs) than the estimated ‘‘counterfactual means’’, such as the marginal rate under control. We can view this empirical example as one where Assumption 2.3 approximately holds in practice, i.e. the treatment effect appears to generalize across observational and RCT populations, but the counterfactual means do not.

## C IHDP Experiment Details

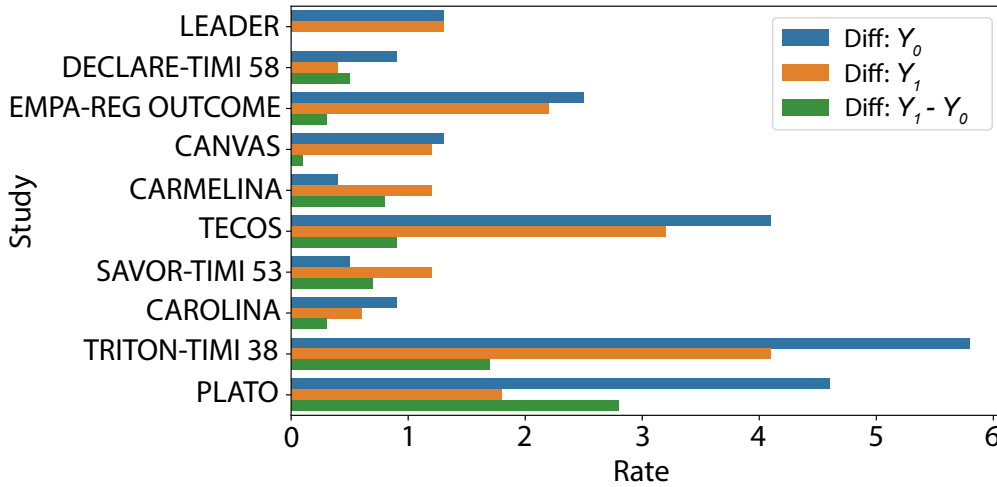
For both the semi-synthetic and real-world experiments, we follow closely the setup proposed by Hussain et al. (2022), with a few differences highlighted below.

### C.1 Confounder Generation & Outcome Simulation

We generate one RCT and one observational study in each of our 100 simulations, with the randomness appearing in our confounder generation, simulation of the potential outcomes, and the amount of noise in each. In the RCT, we retain the original IHDP data, i.e. the covariates and the binary treatment variable, but resample the dataset with equal probability to



(a)



(b)

**Figure 3:** (a) For each binary group indicator  $I$  in the SPRINT Trial, we compare the absolute difference between  $\mathbb{E}[Y_0 | I = 1]$  versus  $\mathbb{E}[Y_0 | I = 0]$ , and similarly for  $Y_1$  and  $Y_1 - Y_0$ , where  $Y$  is a binary variable indicating the observation of the primary composite outcome. The data supporting this plot is taken from Figure 4 of SPRINT Research Group (2015). Generally, the latter difference is smaller (sometimes by an order of magnitude) than the differences for individual potential outcomes. (b) For each attempted replication of an RCT by an Observational study, we compare the differences in the reported incidence rates under treatment, under the control/comparator, and the difference between the two (analogous to the treatment effect). The latter tends to be smaller than both of the differences in counterfactual means in 6 / 10 replications, and smaller than at least one of the differences in counterfactual means in all 10 replications. This data is taken from Table 2 of Franklin et al. (2021), where we use the reported statistics in Table 2.

generate a final dataset of size,  $n_0 = 2955$ . For generation of the observational dataset, we first resample the rows of the IHDP dataset to the desired sample size,  $n = s \cdot n_0$ , but do the resampling in a weighted fashion, such that male infants, infants whose mothers smoked, and infants with working mothers are less prevalent. The weights are set as,

$$w = \frac{1}{1 + \exp(-0.2(1(\text{male infant}) + 1(\text{mother smoked}) + 1(\text{mother worked during pregnancy})))}$$

Note that this differs from the reweighting scheme used in Hussain et al. (2022) in that we use a non-linearity in the reweighting, since we wish for the covariates used in the reweighting (i.e. sex, smoking status, working status) to be effect modifiers.

Next, we generate confounders for the observational dataset. Each confounder,  $z$ , is a function of a subset of the covariates,

$X_s$ , and the treatment,  $A$ :

$$z = X_s^\top \xi + X_s^\top \delta \odot A + \mathcal{N}(0, 1),$$

where  $X_s \in \mathbb{R}^4$  and the coefficients  $\xi$  and  $\delta$  are set as:  $\xi = (0.1, -0.1, 0.2, -0.3, 0.4)$  and  $\delta = (1., -.1, .5, -3, 4)$ .  $X_s$  consists of the following covariates — (“neonatal health index”, “birth order of infant”, “drinks alcohol or not”, “mother finished high school”). Confounder generation for the RCT is similar but does not include any dependency on the treatment:  $X_s^\top \xi + \mathcal{N}(0, 1)$ . We repeat this procedure  $m$  times to yield  $m$  confounders.

To detail the outcome simulation, we borrow notation from Hussain et al. (2022), where we let  $Z \in \mathbb{R}^m$  denote the generated confounder vector and  $X \in \mathbb{R}^{m_x}$  denote the covariate vector, where  $m_x = 28$  is the number of covariates in the original IHDP dataset. Similarly, we let  $\tilde{X} = (A, X^\top)^\top$ . Then, we set the following counterfactual outcome distributions:

$$Y_0 \sim \mathcal{N}\left(\left(\tilde{X} + \frac{1}{2}\mathbf{1}\right)^\top \beta + Z^\top \gamma, 1\right)$$

$$Y_1 \sim \mathcal{N}(\tilde{X}^\top \beta + Z^\top \delta + \omega, 1),$$

where  $\mathbf{1} \in \mathbb{R}^{m_x+1}$  is a vector of ones,  $\beta \in \mathbb{R}^{m_x+1}$  is a vector where each element is randomly sampled from  $(0, 0.1, 0.2, 0.3, 0.4)$  with probabilities  $(0.6, 0.1, 0.1, 0.1, 0.1)$ , and  $\gamma \in \mathbb{R}^m$  is a vector where each element is randomly sampled from one of two vectors with uniform probability depending on the strength of confounding desired:  $(0.1, 0.2, .5, .75, 1.)$  or  $(1., 1.75, 2., 2.25, 2.75)$ . In Figure 1a, for example, we sample from the first vector to generate the confounders, and in Figure 1b, we sample from the second vector. The observed outcome is then set as,  $Y := AY_1 + (1 - A)Y_0$ . Finally,  $\omega = 23$  to bound the magnitude of the counterfactual outcome under treatment. We conceal confounders in order to simulate unobserved confounding, letting  $c_z$  be the number of confounders concealed. As alluded to in the main paper, the order of how we conceal confounders is determined by their “confounding strength”, i.e. from highest to lowest weighted.

## D Women’s Health Initiative (WHI) Experiment Details

We follow substantially the same setup as in Hussain et al. (2022), which we recounted partially in the main paper (Section 5) but do so fully in this section. The WHI conducted several clinical trials as well as an observational study in parallel to study the effect of various hormonal and dietary interventions on the health and quality of life of postmenopausal women. As mentioned in the main paper, of the three clinical trials run by WHI, we use the Postmenopausal Hormone Therapy (PHT) trial for our analysis, which looked at the effect of combination hormone therapy on postmenopausal women aged 50-79 years who had not undergone a hysterectomy (Rossouw et al., 2002). The data used in our analysis is publicly available on BIOLINCC ([https://biolincc.nhlbi.nih.gov/studies/whi\\_ctos](https://biolincc.nhlbi.nih.gov/studies/whi_ctos)).

### D.1 Data

We briefly review the core characteristics of both the RCT and observational study components of the WHI study. The RCT studies the effect of a combination of 2.5mg of medoxyprogesterone and 0.625mg of estrogen on a population of  $N_{HT} = 16608$  postmenopausal women. Each patient is randomly assigned to either the treatment group (i.e. estrogen + progesterone combination is given) or the control group, in which the placebo is given. The outcomes tracked in the RCT are of three categories: 1) cardiovascular events, including coronary heart disease, 2) cancers (endometrial, breast, etc.), and 3) fractures (e.g. hip, bone, etc.). The observational study component studies similar outcomes in a cohort of  $N = 93676$  women. Women were recruited for this component in 1996, and follow-up was done until 2005, which is a similar timeframe as the RCT. Information about therapies that the patients were taking across the follow-up were tracked via questionnaires, which were taken on a yearly basis.

### D.2 Outcome and Intervention

As mentioned in Section 5 of the main paper, we define a binary outcome based on the “global index” score given to each patient, which is a composite index derived from whether or not a patient experiences any one of the following events: coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes. Furthermore, we let  $Y = 1$  if any one of these events is observed in the first seven years of follow-up and  $Y = 0$  otherwise. Notably,  $Y = 0$  may also occur due to censoring.

In terms of the intervention, the RCT is run as an “intention-to-treat” trial. For the observational study component, we determine treatment and control groups based on explicit affirmation or denial of the use of estrogen and progesterone combination therapy in the first three years, which we glean from the annual survey data. Using this procedure, we end up

with a total of  $N_{OS} = 33511$  patients. Finally, we restrict the set of covariates used to those that are measured in both the RCT and the observational study. Each covariate indicates the same meaning, since the same set of questionnaires are used to gather them. The resulting number of covariates is 1576.

### D.3 Experimental Workflow

We detail our experimental setup in this section. We note the following algorithm applies to one row of Table 1 in the main paper. Indeed, to get the remaining results, we re-apply this algorithm after “introducing” some selection bias into the observational dataset. We have the following experimental workflow:

- Step 1: Generate  $B$  bootstrapped datasets of the base WHI observational dataset.
- Step 2: Set list of  $r$  covariate pairs  $X_1, \dots, X_r$ . We use the same set of covariates as used by Hussain et al. (2022), which can be found in Appendix E of their paper, to generate the  $r$  covariate pairs.
- Step 3: For  $i = 1 \rightarrow B$ 
  - Apply **MMR-Contrast** (see Appendix E for implementation details). Set  $\Lambda_{MMR}[i] = 1$  if p-value is  $< 0.05$ , else let  $\Lambda_{MMR}[i] = 0$ .
  - Apply **ATE**. We use the same estimator as Hussain et al. (2022), but average over the entire population to get the ATE from the observational study and RCT, respectively (i.e. set each patient to be part of the same group). Set  $\Lambda_{ATE}[i] = 1$  if the test rejects the null hypothesis and  $\Lambda_{ATE}[i] = 0$  otherwise.
  - For  $j = 1 \rightarrow r$ 
    - Apply **GATE** using the four subgroups derived from  $X_j$ , as in Hussain et al. (2022). Set  $\Lambda_{GATE}[i][j] = 1$  if the test rejects the null hypothesis and  $\Lambda_{GATE}[i][j] = 0$  otherwise.

Thus, the rejection rates for **MMR-Contrast**, **ATE**, **GATE** are  $\frac{1}{B} \sum_i \Lambda_{MMR}[i]$ ,  $\frac{1}{B} \sum_i \Lambda_{ATE}[i]$ ,  $\frac{1}{B \cdot r} \sum_i \sum_j \Lambda_{GATE}[i][j]$ , respectively. We repeat the above workflow to the WHI dataset that has induced selection bias. To add selection bias to the data, with probability  $p$ , we drop patients who were not exposed to the intervention and did not experience the event. To obtain the results for Table 1, we run our experimental procedure for  $p = (0., 0.05, 0.10, 0.15)$ .

## E Details on Implementation of the MMR-Contrast Method

The implementation of the MMR-Contrast method references the workflow illustrated in (Hussain et al., 2022) and (Muandet et al., 2020): the former for signal calculation and the latter for significance testing.

### E.1 Calculation of signal difference

As elaborated in the main text, in the combined data (combining the RCT and observational study), for each observation  $i$  producing data  $(y_i, s_i, a_i, x_i)$ , we define the true signal difference as,

$$\psi_i = \frac{1}{P(S=0|X=x_i)} \left\{ \mathbf{1}(s_i=0) \left[ [\mu_1(x_i) - \mu_0(x_i)] - \left[ \frac{\mathbf{1}(a_i=1)}{P(A=1|S=0)} - \frac{\mathbf{1}(a_i=0)}{P(A=0|S=0)} \right] y_i \right] + \right. \\ \left. \mathbf{1}(s_i=1) \frac{P(S=0|X=x_i)}{P(S=1|X=x_i)} \left[ \frac{\mathbf{1}(a_i=1)(y_i - \mu_1(x_i))}{P(A=1|S=1, X=x_i)} - \frac{\mathbf{1}(a_i=0)(y_i - \mu_0(x_i))}{P(A=0|S=1, X=x_i)} \right] \right\}$$

where  $\mu_1(x_i) = \mathbb{E}[Y|S=0, A=1, X=x_i]$  and  $\mu_0(x_i) = \mathbb{E}[Y|S=0, A=0, X=x_i]$ . Note that the true signal difference includes several unknown nuisance functions that need to be estimated:

- Response surface:  $\mu_1(X), \mu_0(X)$
- Selection propensity:  $P(S=1|X), P(S=0|X)$
- Treatment propensity in the observational study:  $P(A=1|S=1, X), P(A=0|S=1, X)$
- Treatment propensity in the RCT:  $P(A=1|S=0), P(A=0|S=0)$

The treatment propensity in the RCT is estimated with the empirical probability of treatment within the RCT data. The response surface, selection propensity and treatment propensity in the observational study are estimated using cross-fitting: the combined data is randomly split into  $K=3$  folds, and the nuisance functions used in each fold are estimated with data out of that fold, using the following models with grid search for hyperparameters. Default hyperparameters in *scikit-learn* for the linear regression model were used. The best hyperparameters found for the gradient boosting classifier, also in *scikit-learn*,

were as follows: “learning-rate”: 0.01, “n-estimators”: 50, “max-depth”: 2, “min-samples-leaf”: 50, “min-samples-split”: 50, “max-features”: “sqrt” (Pedregosa et al., 2011).

	Response surface	Selection propensity	Treatment propensity (observational)
IDHP	Linear regression	Gradient boosting classifier	Gradient boosting classifier
WHI	Gradient boosting classifier	Gradient boosting classifier	Gradient boosting classifier

As an aside, in Figure 2(a) where we compare the performance of statistics using estimated signals and true signals, we plug in the response surface and selection propensity model implied by our simulation settings into  $\psi_i$  to get the true signal difference.

## E.2 Hypothesis testing

After obtaining the estimated signal difference  $\hat{\psi}_i$  by plugging in the estimated nuisance functions into  $\psi_i$ , the test statistic is calculated as,

$$n\hat{\mathbb{M}}_n^2 = \frac{1}{(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \hat{\psi}_i k(x_i, x_j) \hat{\psi}_j$$

where  $k(\cdot, \cdot)$  is set as a polynomial kernel of order 3. One may also use a laplacian kernel or RBF kernel, although we found the polynomial and laplacian kernels to work best in practice. To obtain the  $p$ -value for the test, we follow (Muandet et al., 2020) and generate  $B = 100$  samples of multinomials  $\mathbf{w}_k = (w_{k1}, w_{k2}, \dots, w_{kn})^\top \sim \text{Multinom}(n, (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}))$ ,  $k = 1, 2, \dots, B$ . For each  $k$ , we define the bootstrap sample of the null distribution:

$$n\hat{\mathbb{M}}_{n(k)}^2 = n \sum_{i,j \in \mathcal{I}, i \neq j} \frac{w_{ki} - 1}{n} \hat{\psi}_i k(x_i, x_j) \hat{\psi}_j \frac{w_{kj} - 1}{n}$$

The  $p$ -value is then calculated as

$$\frac{\left[ \sum_{k=1}^B \mathbf{1}(n\hat{\mathbb{M}}_n^2 \leq n\hat{\mathbb{M}}_{n(k)}^2) \right] + 1}{B + 1}$$

Note that we do not re-estimate the propensity score function in each bootstrap iteration.

## F Beyond Testing CATE Signals

In this section, we provide a different formulation of our falsification procedure that tests the potential outcome signals for  $\mathbb{E}[Y_a|X]$ ,  $a \in \{0, 1\}$ , individually, instead of the signal for the contrast  $\mathbb{E}[Y_1 - Y_0|X]$ . This demonstrates that our formulation can be adapted to testing other functions of the potential outcome distribution, other than the one we originally considered.

First, we modify our external validity assumption to accommodate testing individual potential outcomes:

**Assumption F.1** (*External Validity: Observational Study to RCT Transportability of Potential Outcomes*). We assume the following:

- *Mean Exchangeability* —  $\mathbb{E}[Y_a|X = x] = \mathbb{E}[Y_a|X = x, S = s]$ ,  $\forall x \in \mathcal{X}$ ,  $\forall s \in \{0, 1\}$ , and  $\forall a \in \{0, 1\}$ .
- *Positivity of Selection* —  $\mathbb{P}(X = x|S = 0) > 0 \implies \mathbb{P}(X = x|S = 1) > 0$ ,  $\forall x \in \mathcal{X}$ .

Now, we will introduce additional notation for our signal functions. Namely, we have the outcome signal from the RCT as follows,

$$\begin{aligned} \psi_0^a &= \frac{\mathbf{1}\{S = 0\}}{P(S = 0|X)} \cdot \frac{Y\mathbf{1}\{A = a\}}{P(A = a|S = 0)}, a \in \{0, 1\} \\ \boldsymbol{\psi}_0 &= (\psi_0^0, \psi_0^1)^\top \end{aligned} \tag{14}$$

Similarly, we have the following outcome signal in the RCT population, but estimated from observational data, as developed in the main paper,

$$\begin{aligned} \psi_1^a &= \frac{1}{P(S = 0|X)} \left[ \mathbf{1}\{S = 0\} \mu_a(X) + \mathbf{1}\{S = 1\} \frac{P(S = 0|X)}{P(S = 1|X)} \frac{\mathbf{1}\{A = a\}(Y - \mu_a(X))}{P(A = a|S = 1, X)} \right], a \in \{0, 1\} \\ \boldsymbol{\psi}_1 &= (\psi_1^0, \psi_1^1)^\top \end{aligned} \tag{15}$$

Note that the main difference here compared to the main paper is that we define signal functions individually for each potential outcome and then let  $\psi_0$  and  $\psi_1$  be a vector of signals. Now, we have the following proposition, which shows that the vector signals are unbiased for the potential outcomes in the RCT population:

**Proposition F.1** (*Potential Outcome Signals from the RCT and Observational Data*). *Under Assumption 2.2 (internal validity of the RCT), the instance-wise potential outcome vector  $\psi_0$  in Eq. (14), which uses the outcome information from the RCT, is unbiased, i.e.,  $\mathbb{E}[\psi_0|X] = \mathbb{E}[\mathbf{Y}|X, S = 0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$ . Furthermore, under Assumption 2.1 and Assumption F.1, the instance-wise potential outcome vector  $\psi_1$  in Eq. (15), which uses the outcome information from the observational data, is unbiased for the potential outcomes in the RCT population, i.e.  $\mathbb{E}[\psi_1|X] = \mathbb{E}[\mathbf{Y}|X, S = 0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$ .*

*Proof.* We first show  $\mathbb{E}[\psi_0|X] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$ , i.e.  $\mathbb{E}[\psi_0^a|X] = \mathbb{E}[Y_a|X, S = 0]$ ,  $a \in \{0, 1\}$

$$\begin{aligned}
 \mathbb{E}[\psi_0^a|X] &= \mathbb{E}\left[\frac{\mathbf{1}\{S = 0\}}{P(S = 0|X)} \frac{Y \mathbf{1}\{A = a\}}{P(A = a|S = 0)} \middle| X\right] \\
 &= \frac{1}{P(S = 0|X)P(A = a|S = 0)} \mathbb{E}[\mathbf{1}\{S = 0, A = a\}Y|X] \\
 &= \frac{1}{P(S = 0|X)P(A = a|S = 0, X)} \mathbb{E}[\mathbf{1}\{S = 0, A = a\}Y|X] \\
 &= \frac{P(S = 0, A = a|X)}{P(S = 0|X)P(A = a|S = 0, X)} \mathbb{E}[Y|X, S = 0, A = a] \\
 &= \mathbb{E}[Y|X, S = 0, A = a] \\
 &= \mathbb{E}[Y_a|X, S = 0],
 \end{aligned} \tag{16}$$

where (16) is from fixed probability of assignment in Assumption 2.2, and (17) is from consistency and ignorability in Assumption 2.2.

We then show  $\mathbb{E}[\psi_1|X] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 1]$ , i.e.  $\mathbb{E}[\psi_1^a|X] = \mathbb{E}[Y_a|X, S = 1]$ ,  $a \in \{0, 1\}$

$$\begin{aligned}
 \mathbb{E}[\psi_1^a|X] &= \mathbb{E}\left[\frac{1}{P(S = 0|X)} \left[ \mathbf{1}\{S = 0\} \mu_a(X) + \mathbf{1}\{S = 1\} \frac{P(S = 0|X)}{P(S = 1|X)} \frac{\mathbf{1}\{A = a\}(Y - \mu_a(X))}{P(A = a|S = 1, X)} \right] \middle| X\right] \\
 &= \frac{\mathbb{E}[\mathbf{1}\{S = 0\}|X] \mu_a(X)}{P(S = 0|X)} + \frac{\mathbb{E}[\mathbf{1}\{S = 1, A = a\}(Y - \mu_a(X))|X]}{P(S = 1|X)P(A = a|S = 1, X)} \\
 &= \frac{P(S = 0|X) \mu_a(X)}{P(S = 0|X)} + \frac{P(S = 1, A = a|X) \mathbb{E}[(Y - \mu_a(X))|X, S = 1, A = a]}{P(S = 1|X)P(A = a|S = 1, X)} \\
 &= \mu_a(X) + \mathbb{E}[(Y - \mu_a(X))|X, S = 1, A = a] \\
 &= \mu_a(X) + \mathbb{E}[Y|X, S = 1, A = a] - \mu_a(X) \\
 &= \mu_a(X) + \mathbb{E}[Y_a|X, S = 1] - \mu_a(X) \\
 &= \mathbb{E}[Y_a|X, S = 1],
 \end{aligned} \tag{18}$$

where (18) is from consistency and ignorability in Assumption 2.1. The proposition is now proven.  $\square$

We can show a similar corollary to Corollary 3.1 in the main paper, where we developed the null hypothesis on the CATE signals. Now, we do so for the potential outcome vector signals. Namely, we have,

**Corollary F.1** (*Null Hypothesis on Potential Outcome Difference*). *Define  $\psi = \psi_1 - \psi_0$  as the instance-wise signal difference between the observational and RCT potential outcome estimates. Then, under the null hypothesis, i.e. under Assumptions 2.1 and 2.2 and Assumption F.1, we have it that  $\mathbb{E}[\psi|X] = \mathbf{0}$ .*

*Proof.* If Assumptions 2.1 and 2.2 and Assumption F.1 hold, then Proposition F.1 implies that  $\mathbb{E}[\psi_0|X] = \mathbb{E}[\psi_1|X] = \mathbb{E}[\mathbf{Y}|X, S = 0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$ .  $\square$

Our assumptions, i.e. Assumption 2.2, Assumption 2.1, and Assumption F.1, give us a set of conditional moment restrictions (CMRs) on the signal difference,  $\psi$ , which is a difference of vector signals:

$$H_0 : \mathbb{E}[\psi|X] = \mathbf{0} \quad P_X\text{-almost surely} \tag{19}$$

As before,  $P_X$  is the distribution of  $X$  on the joint distribution of the RCT and observational study. By the law of iterated expectations, akin to the development in Proposition 3.3, Eq. (19) implies an infinite set of unconditional moment restrictions,

$$\mathbb{E}[\boldsymbol{\psi}^\top \mathbf{f}(X)] = 0, \forall \mathbf{f} \in \mathcal{F} \times \mathcal{F}, \quad (20)$$

where  $\mathcal{F}$  is the set of measurable functions on  $\mathcal{X}$ . Note that now,  $\mathbf{f}$  is a vector-valued function, where  $\mathbf{f}(X) = (f_0(X), f_1(X))^\top$ . Now, as in the main paper, we follow the CMR testing procedure presented in Muandet et al. (2020), where we let  $\mathcal{F}$  be a RKHS and use the maximum moment restriction (MMR) within the unit ball of the RKHS as our test statistic. Following this, we present the following theorem, which is a modified version of Theorem 3.1.

**Theorem F.1** (*Maximum Moment Restriction-based test for Potential Outcomes*). *Let  $\mathcal{F}$  be a RKHS with reproducing kernel  $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  that is ISPD, continuous and bounded, equipped with inner product  $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ . Denote  $\mathcal{F}^2$  as the product RKHS  $\mathcal{F} \times \mathcal{F}$  equipped with inner product  $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2} = \langle (f_1, f_2)^\top, (g_1, g_2)^\top \rangle_{\mathcal{F}^2} := \langle f_1, g_1 \rangle_{\mathcal{F}} + \langle f_2, g_2 \rangle_{\mathcal{F}}$ . Suppose the elements of  $|\mathbb{E}[\boldsymbol{\psi}|X]| < \infty$  almost surely in  $P_X$ , and  $\mathbb{E}[[k(X, X')\boldsymbol{\psi}^\top \boldsymbol{\psi}']^2] < \infty$  where  $(\boldsymbol{\psi}', X')$  is an independent copy of  $(\boldsymbol{\psi}, X)$ . Let  $\mathbb{M}^2 = \sup_{\mathbf{f} \in \mathcal{F}^2, \|\mathbf{f}\| \leq 1} (\mathbb{E}[\boldsymbol{\psi}^\top \mathbf{f}(X)])^2$ . Then,*

1. *The conditional moment testing problem in Eq. 19 can be reformulated in terms of the MMR as  $H'_0 : \mathbb{M}^2 = 0$ ,  $H'_1 : \mathbb{M}^2 \neq 0$ .*

Further, let the test statistic be the empirical estimate of  $\mathbb{M}^2$ ,

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} k(x_i, x_j) \boldsymbol{\psi}_i^\top \boldsymbol{\psi}_j$$

2. *Then, under  $H'_0$ ,*

$$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

where  $Z_j$  are independent standard normal variables and  $\lambda_j$  are the eigenvalues for  $k(x, x')\boldsymbol{\psi}^\top \boldsymbol{\psi}'$ .

3. *Under  $H'_1$ ,*

$$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$

where  $\sigma^2 = \text{var}_{(\boldsymbol{\psi}, X)}[\mathbb{E}_{(\boldsymbol{\psi}', X')}[k(X, X')\boldsymbol{\psi}^\top \boldsymbol{\psi}']]$

*Proof.* The proof is very similar to how we proved Theorem 3.1. Let us define the following operator,

$$M\mathbf{f} = \mathbb{E}[\boldsymbol{\psi}^\top \mathbf{f}(X)] \quad (21)$$

where  $\mathbf{f} \in \mathcal{F}^2$ . Since the elements of  $|\mathbb{E}[\boldsymbol{\psi}|X]| < \infty$  almost surely in  $P_X$ ,  $M$  is a bounded linear operator. By Riesz representation theorem, there exists a unique  $\mathbf{g} \in \mathcal{F}^2$  such that

$$M\mathbf{f} = \langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2}$$

where

$$\mathbf{g} = \mathbb{E}[\boldsymbol{\psi}k(X, \cdot)].$$

Therefore, it follows that

$$\mathbb{M}^2 = \sup_{\mathbf{f} \in \mathcal{F}^2, \|\mathbf{f}\| \leq 1} \left( \mathbb{E}[\boldsymbol{\psi}^\top \mathbf{f}(X)] \right)^2 = \sup_{\mathbf{f} \in \mathcal{F}^2, \|\mathbf{f}\| \leq 1} \langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2}^2 = \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|}, \mathbf{g} \right\rangle_{\mathcal{F}^2}^2 = \|\mathbf{g}\|^2$$

Since  $\mathbb{M}^2 = \|\mathbf{g}\|^2$ , the first statement in Theorem F.1 is essentially

$$\mathbb{E}[\boldsymbol{\psi}|X] = \mathbf{0}, P_X\text{-almost surely} \Leftrightarrow \|\mathbf{g}\|^2 = 0$$

That is,  $\mathbf{g} \in \mathcal{F}^2$  fully captures the information of the CMR for all  $x \in \mathcal{X}$ . This equivalence, which we will now prove, is crucial since our statistical test is based on  $\|\mathbf{g}\|^2$  and its estimates, while Corollary F.1 is directed to the CMR:



( $\Rightarrow$ ) We note that since  $\mathcal{F}^2$  is a Hilbert space, it follows that  $\mathbf{g} \in \mathcal{F}^2$ , and from (20),  $\forall \mathbf{f} \in \mathcal{F}^2$ ,  $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2} = \mathbb{E}[\boldsymbol{\psi}^\top \mathbf{f}(X)] = 0$ .  $\mathbf{g}$  can now only be a zero vector. Therefore,  $\|\mathbf{g}\|^2 = 0$ .

( $\Leftarrow$ )

$$\begin{aligned}
 & \|\mathbf{g}\|^2 = 0 \\
 \Rightarrow & \|\mathbb{E}[\boldsymbol{\psi}k(X, \cdot)]\|^2 = 0 \\
 \Rightarrow & \|\mathbb{E}[\mathbb{E}[\boldsymbol{\psi}|X]k(X, \cdot)]\|^2 = 0 \\
 \Rightarrow & \left\| \int_{\mathcal{X}} k(x, \cdot) \mathbb{E}[\boldsymbol{\psi}|x] p_X(x) dx \right\|^2 = 0 \\
 \Rightarrow & \iint_{\mathcal{X} \times \mathcal{X}} p_X(x) \mathbb{E}[\boldsymbol{\psi}^\top |x] k(x, x') \mathbb{E}[\boldsymbol{\psi}|x'] p_X(x') dx dx' = 0 \\
 \Rightarrow & \|\mathbb{E}[\boldsymbol{\psi}|x] p_X(x)\|^2 = 0 \quad (\because k(\cdot, \cdot) \text{ is ISPD}) \\
 \Rightarrow & \mathbb{E}[\boldsymbol{\psi}|x] = \mathbf{0}, P_X\text{-almost surely}
 \end{aligned}$$

Finally, we move to the second and third statements of Theorem F.1, which define the estimator and its statistical properties. Since  $\mathbb{M}^2 = \|\mathbf{g}\|^2 = \|\mathbb{E}[\boldsymbol{\psi}k(X, \cdot)]\|^2 = \mathbb{E}[\mathbb{E}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']]$  where  $(X, \boldsymbol{\psi})$  and  $(X', \boldsymbol{\psi}')$  are independently and identically distributed, we may use a  $U$ -statistic to estimate  $\mathbb{M}^2$ , which is exactly

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \boldsymbol{\psi}_i^\top k(x_i, x_j) \boldsymbol{\psi}_j$$

Now note that in Lemma A.1, if we set  $W = (\boldsymbol{\psi}, X)$ ,  $h(W, W') = \boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}'$ ,  $\theta = \mathbb{M}^2$ ,  $\zeta_1 = \sigma^2$ , the second and third statements of Theorem 3.1 holds as long as  $\mathbb{M}^2 = 0 \Leftrightarrow \sigma^2 = \text{var}_{(\boldsymbol{\psi}, X)}[\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']] = 0$ , which we will now show:

( $\Rightarrow$ )

$$\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}'] = \langle \boldsymbol{\psi}k(X, \cdot), \mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}'k(X', \cdot)] \rangle_{\mathcal{F}^2} = \|\boldsymbol{\psi}k(X, \cdot)\| \left\langle \frac{\boldsymbol{\psi}k(X, \cdot)}{\|\boldsymbol{\psi}k(X, \cdot)\|}, \mathbf{g} \right\rangle_{\mathcal{F}^2}$$

Now since

$$\frac{\boldsymbol{\psi}k(X, \cdot)}{\|\boldsymbol{\psi}k(X, \cdot)\|} \in \mathcal{F}^2, \left\| \frac{\boldsymbol{\psi}k(X, \cdot)}{\|\boldsymbol{\psi}k(X, \cdot)\|} \right\| = 1$$

and

$$\mathbb{M}^2 = 0 \Rightarrow \sup_{\mathbf{f} \in \mathcal{F}^2, \|\mathbf{f}\| \leq 1} \langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2} = 0 \Rightarrow \langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^2} = 0, \forall \mathbf{f} \in \mathcal{F}^2, \|\mathbf{f}\| \leq 1$$

We conclude

$$\mathbb{M}^2 = 0 \Rightarrow \left\langle \frac{\boldsymbol{\psi}k(X, \cdot)}{\|\boldsymbol{\psi}k(X, \cdot)\|}, \mathbf{g} \right\rangle_{\mathcal{F}^2} = 0 \Rightarrow \mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}'] = 0 \Rightarrow \text{var}_{(\boldsymbol{\psi}, X)}[\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']] = 0$$

( $\Leftarrow$ )

We first note that  $\text{var}_{(\boldsymbol{\psi}, X)}(\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']) = 0$  implies that  $\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']$  is a constant  $P_{(\boldsymbol{\psi}, X)}$ -almost surely. We denote this constant as  $c$  so we have

$$\mathbb{E}_{(\boldsymbol{\psi}', X')}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}'] = c, P_{(\boldsymbol{\psi}, X)}\text{-almost surely} \tag{22}$$

From the definition of  $\psi$ , let  $X = x^*$  be in the support of the observational study, then

$$\begin{aligned} \mathbb{E}[\psi|S = 1, X = x^*] &= \frac{1}{P(S = 1|X = x^*)} \mathbb{E} \left[ \left( \frac{\mathbf{1}(A = 1)(Y - \mu_1(x^*))}{P(A = 1|S = 1, X = x^*)}, \right. \right. \\ &\quad \left. \left. \frac{\mathbf{1}(A = 0)(Y - \mu_0(x^*))}{P(A = 0|S = 1, X = x^*)} \right)^\top \middle| S = 1, X = x^* \right] \\ &= \frac{1}{P(S = 1|X = x^*)} (\mathbb{E}[Y - \mu_1(x^*)|A = 1, S = 1, X = x^*], \\ &\quad \mathbb{E}[Y - \mu_0(x^*)|A = 0, S = 1, X = x^*])^\top \\ &= \mathbf{0}, \end{aligned}$$

where the last equality stems from the definition of  $\mu_1$  and  $\mu_0$ . Now note that

$$\begin{aligned} \mathbb{E}_\psi[\mathbb{E}_{(\psi', X')}[\psi^\top k(X, X')\psi'|S = 1, X = x^*]] &= \mathbb{E}_\psi[\mathbb{E}_{(\psi', X')}[\psi^\top k(x^*, X')\psi'|S = 1, X = x^*]] \\ &= (\mathbb{E}_\psi[\psi|S = 1, X = x^*])^\top \mathbb{E}_{(\psi', X')}[k(x^*, X')\psi'] \\ &= \mathbf{0}^\top \mathbb{E}_{(\psi', X')}[k(x^*, X')\psi'] = 0 \end{aligned}$$

But also we have, from (22),

$$\mathbb{E}_\psi[\mathbb{E}_{(\psi', X')}[\psi^\top k(X, X')\psi'|S = 1, X = x^*]] = \mathbb{E}_\psi[c] = c$$

Therefore, we have  $c = 0$  and thus

$$\mathbb{M}^2 = \mathbb{E}_{(\psi, X)}[\mathbb{E}_{(\psi', X')}[\psi^\top k(X, X')\psi']] = \mathbb{E}_{(\psi, X)}[0] = 0$$

so this side of the arrow is also proven. □

We label this alternate formulation, where we test on the potential outcomes directly instead of the contrast, as **MMR-Absolute**. We give the rejection rate of **MMR-Absolute** under different amounts of selection bias induced in the WHI dataset in Table 2. We find that the **MMR-Absolute** approach vastly over-rejects, indicating the utility of testing the causal contrast as opposed to the absolute potential outcomes.

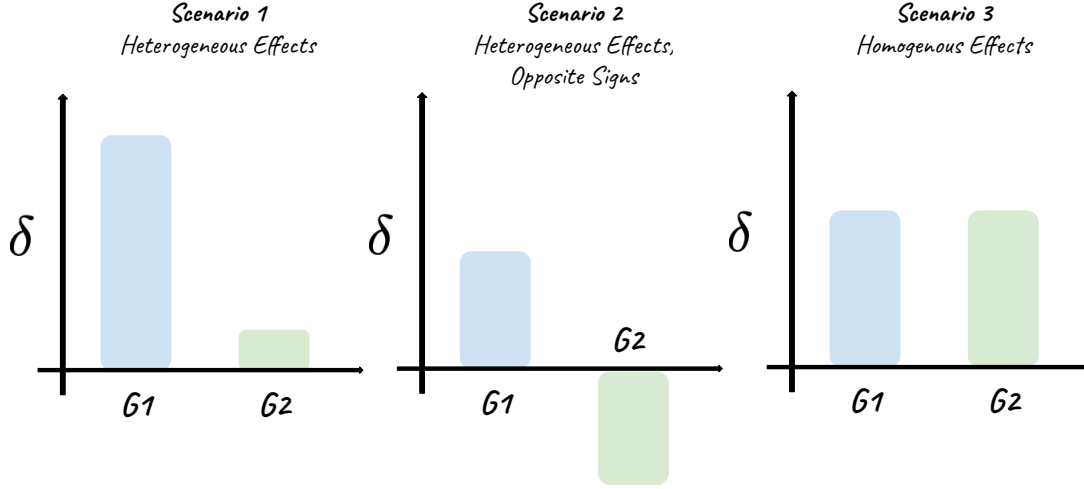
<i>Selection Bias</i>	<b>MMR-Contrast</b>	<b>MMR-Absolute</b>	ATE	GATE
$p = 0$	0.29	1.0	0.32	<b>0.17</b>
$p = 0.05$	<b>0.67</b>	1.0	0.58	0.40
$p = 0.10$	<b>0.94</b>	1.0	0.88	0.67
$p = 0.15$	<b>1.0</b>	1.0	0.98	0.91

**Table 2:** Rejection rate when introducing different amounts of selection bias into the observational data in WHI study.  $p$  stands for the strength of selection introduced in the the data (refer to Section 5 for details).

## G When does testing for bias across subgroups improve power?

In our experimental results, we find that the GATE approach has limited power compared to the ATE approach. Indeed, the performance of GATE versus ATE depends in part on the choice of subgroups used for GATE. In the extreme, if the difference in effect is identical across all subgroups, testing for differences in ATE may have higher power once multiple-testing corrections are applied. To build intuition, we will provide a simple example for when a GATE-based test might have higher power compared to an ATE-based test. We will then formalize this example and provably show under what conditions a GATE-based test would have higher asymptotic power compared to an ATE-based test. When referring to the test that tests differences of GATEs or ATEs, we will use the bold form: **GATE** and **ATE**. When referring to the causal quantity itself, we will simply use GATE and ATE.

**Toy Example:** To build intuition, we will use a toy example to construct three scenarios in which the asymptotic power between **GATE** and **ATE** may differ. Consider testing whether there is bias in a population, where the null hypothesis is that the population mean is zero. Let there be two subgroups in the population, **G1** and **G2**. Finally, let  $\delta$  be a term denoting the asymptotic bias. Figure 4 shows three separate scenarios:



**Figure 4:** Barplot depiction of three toy scenarios, where we plot the asymptotic bias, denoted by  $\delta$  (see Equations (31) and (32)), of the observational estimator in each subgroup. Our goal is to detect, from finite samples, whether or not this asymptotic bias is non-zero for any subgroup. In scenario 3, pooling the data and testing for the overall bias (the **ATE** approach) yields better power than testing for differences across subgroups. Explicitly testing the bias in each subgroup (the **GATE** approach) is beneficial in scenarios like 1 and 2 where heterogeneity exists. The x-axis contains the group name, and the y-axis indicates the magnitude of  $\delta$ .

- In scenario 1, the bias in **G1** is significantly higher than the bias in **G2**. As we formalize below, **GATE** will have higher power than **ATE** as  $|\delta|$  gets larger and the sample size of **G1** is reasonable. See below for precise conditions.
- In scenario 2, the bias in the two subgroups have the same magnitude but are in opposite directions. Below, we show that **GATE** has better power than **ATE** in this scenario, given a large enough  $|\delta|$  to overcome the penalty of multiple hypothesis testing. This result is intuitive since testing differences in ATE would fail to reject the null since the average effect over the entire population would be close to zero.
- In scenario 3, the bias is the same magnitude and direction in both subgroups. We show below that the **ATE** has better power than **GATE** regardless of what the magnitude of  $\delta$  is. Intuitively, pooling together the two subgroups would yield a larger sample to detect the bias.

In the subsequent paragraphs, we will formalize these three scenarios in the context of our setting, where we have estimates from observational and RCT data. Note that the theoretical framework that we introduce below covers these three scenarios as well as others.

### G.1 Notation and Assumptions

We recall some notation and definitions from (Hussain et al., 2022).

**Definition G.1** (GATE, Hussain et al. (2022)). We define the group average treatment effect (GATE) as

$$\tau_i := \mathbb{E}[Y_1 - Y_0 \mid G = i, S = 0] \quad (23)$$

where  $G$  is the group indicator variable taking values  $\{1, 2\}$ , and  $S = 0$  indicates the RCT population.

The GATE estimator for subgroup  $i$  using RCT data will be denoted,  $\hat{\tau}_i(0)$ , while the estimator using observational data will be denoted,  $\hat{\tau}_i(1)$ .

**Definition G.2** (ATE). We define the average treatment effect (ATE) as

$$\tau := \mathbb{E}[Y_1 - Y_0 \mid S = 0] \quad (24)$$

where  $S = 0$  indicates the RCT population.

Akin to the GATE estimators, the ATE estimator using RCT data will be denoted,  $\hat{\tau}(0)$ , while the estimator using observational data will be denoted,  $\hat{\tau}(1)$ . Writing  $\rho_{i0}$  ( $\rho_{i1}$ ) as the proportion of observations in the RCT (the observational study) that belongs to subgroup  $i$ , we then modify Assumption 2.4 from Hussain et al. (2022) as follows, :

**Assumption G.1.** All GATE estimators are pointwise asymptotically normally distributed and independent

$$\sqrt{\rho_{i0}N_0}(\hat{\tau}_i(0) - \tau_i(0))/\hat{\sigma}_i(0) \xrightarrow{d} \mathcal{N}(0, 1) \quad (25)$$

$$\sqrt{\rho_{i1}N_1}(\hat{\tau}_i(1) - \tau_i(1))/\hat{\sigma}_i(1) \xrightarrow{d} \mathcal{N}(0, 1) \quad (26)$$

Here,  $\xrightarrow{d}$  denotes convergence in distribution, and  $\hat{\sigma}_i^2(k)$  is an estimate of the variance that converges in probability to  $\sigma_i^2(k)$ , the asymptotic variance of  $\sqrt{\rho_{ik}N_k}(\hat{\tau}_i(k) - \tau_i(k))$ , for  $k = 0$  and  $k = 1$ .

In addition to assumptions on the GATE estimators, we also have assumptions on the asymptotic distributions of the ATE estimators for both studies:

**Assumption G.2.** Both ATE estimators are asymptotically normally distributed and independent

$$\sqrt{N_0}(\hat{\tau}(0) - \tau(0))/\hat{\sigma}(0) \xrightarrow{d} \mathcal{N}(0, 1) \quad (27)$$

$$\sqrt{N_1}(\hat{\tau}(1) - \tau(1))/\hat{\sigma}(1) \xrightarrow{d} \mathcal{N}(0, 1) \quad (28)$$

where  $\hat{\sigma}^2(k)$  is an estimate of the variance that converges in probability to  $\sigma^2(k)$ , the asymptotic variance of  $\sqrt{N_k}(\hat{\tau}(k) - \tau(k))$ , for  $k = 0$  and  $k = 1$ .

## G.2 Theoretical Example

Given the assumptions and definitions, we present a formal example:

**Example G.1.** Suppose there are two subgroups in the RCT and observational study. To reflect the consistency of the RCT GATE estimators and quantify the bias of the GATE estimators from the observational study, we define

$$(RCT, \text{group 1}) \quad \tau_1(0) = \tau_1 \quad (29)$$

$$(RCT, \text{group 2}) \quad \tau_2(0) = \tau_2 \quad (30)$$

$$(OBS, \text{group 1}) \quad \tau_1(1) = \tau_1 + \delta_1 \quad (31)$$

$$(OBS, \text{group 2}) \quad \tau_2(1) = \tau_2 + \delta_2 \quad (32)$$

For simplicity, we assume that in both the RCT and observational study, half of the population is in group 1 and half of the population is in group 2, i.e.  $\rho_{i0} = \rho_{i1} = 1/2$  for  $i = 1, 2$ . Then we have

$$\tau(0) = \frac{\tau_1(0) + \tau_2(0)}{2} = \frac{\tau_1 + \tau_2}{2} \quad (33)$$

$$\tau(1) = \frac{\tau_1(1) + \tau_2(1)}{2} = \frac{\tau_1 + \delta_1 + \tau_2 + \delta_2}{2} \quad (34)$$

Lastly, we introduce the following shorthand notations, writing the total sample size  $N = N_0 + N_1$  and letting  $N_0 = \rho N$ ,  $N_1 = (1 - \rho)N$ :

$$\sigma = \sqrt{N_0 + N_1} \sqrt{\frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}} = \sqrt{\frac{\sigma^2(0)}{\rho} + \frac{\sigma^2(1)}{1 - \rho}} \quad (35)$$

$$\sigma_1 = \sqrt{\rho_{10}N_0 + \rho_{11}N_1} \sqrt{\frac{\sigma_1^2(0)}{\rho_{10}N_0} + \frac{\sigma_1^2(1)}{\rho_{11}N_1}} = \sqrt{\frac{\sigma_1^2(0)}{\rho} + \frac{\sigma_1^2(1)}{1 - \rho}} \quad (36)$$

$$\sigma_2 = \sqrt{\rho_{20}N_0 + \rho_{21}N_1} \sqrt{\frac{\sigma_2^2(0)}{\rho_{20}N_0} + \frac{\sigma_2^2(1)}{\rho_{21}N_1}} = \sqrt{\frac{\sigma_2^2(0)}{\rho} + \frac{\sigma_2^2(1)}{1 - \rho}} \quad (37)$$

To simplify the development, we will make the following assumption for this example:

**Assumption G.3.** Assume that  $\sigma^2(0) = \sigma_1^2(0) = \sigma_2^2(0)$  and  $\sigma^2(1) = \sigma_1^2(1) = \sigma_2^2(1)$ , so that we can write Equations (35) to (37) as,

$$\sigma = \sigma_1 = \sigma_2 = \sqrt{\frac{\sigma^2(0)}{\rho} + \frac{\sigma^2(1)}{1 - \rho}} \quad (38)$$

The asymptotic power of the **ATE** and **GATE** can then be given by the following propositions:

**Proposition G.1** (Asymptotic power of **ATE**). Under Assumption G.3, the asymptotic power of **ATE** as  $N \rightarrow \infty$  (holding  $\rho$  as constant) is given by

$$1 - \left[ \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} + z_{\alpha/2} \right) - \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} - z_{\alpha/2} \right) \right]$$

*Proof.* From Proposition 2.1 of (Hussain et al., 2022), given the asymptotic distributions from Assumption G.2 and Equations (33), (34), we have  $\tau(1) - \tau(0) = \frac{\delta_1 + \delta_2}{2}$  and thus

$$\frac{\hat{\tau}(1) - \hat{\tau}(0) - \frac{\delta_1 + \delta_2}{2}}{\hat{\sigma}/\sqrt{N}} \xrightarrow{d} \mathcal{N}(0, 1) \quad (39)$$

which allows us to construct a  $Z$ -test on the null hypothesis  $H_0 : \frac{\delta_1 + \delta_2}{2} = 0$  based on the rejection region

$$\left| \frac{\hat{\tau}(1) - \hat{\tau}(0)}{\hat{\sigma}/\sqrt{N}} \right| > z_{\alpha/2}$$

The asymptotic power of the  $Z$ -test under the alternative hypothesis distribution shown in (39) is then, from Theorems 10.4, 10.6 in (Wasserman, 2004)

$$1 - \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} + z_{\alpha/2} \right) + \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} - z_{\alpha/2} \right) = 1 - \left[ \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} + z_{\alpha/2} \right) - \Phi \left( \frac{|\frac{\delta_1 + \delta_2}{2}|}{\sigma/\sqrt{N}} - z_{\alpha/2} \right) \right]$$

□

**Proposition G.2** (Asymptotic power of **GATE**). Under Assumption G.3, the asymptotic power of **GATE** is given by

$$1 - \left[ \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_1|}{\sigma/\sqrt{N}} + z_{\alpha/4} \right) - \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_1|}{\sigma/\sqrt{N}} - z_{\alpha/4} \right) \right] \left[ \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_2|}{\sigma/\sqrt{N}} + z_{\alpha/4} \right) - \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_2|}{\sigma/\sqrt{N}} - z_{\alpha/4} \right) \right]$$

*Proof.* With arguments similar to Proposition G.1, since the total sample size for subgroup  $i$  is  $\rho_{i0}N_0 + \rho_{i1}N_1 = N/2$ , the asymptotic power of the  $Z$ -test comparing the GATE estimates for group  $i$  would be ( $i \in \{1, 2\}$ )

$$\begin{aligned} \xi_i &= 1 - \left[ \Phi \left( \frac{|\delta_i|}{\sigma_i/\sqrt{N/2}} + z_{\alpha/4} \right) - \Phi \left( \frac{|\delta_i|}{\sigma_i/\sqrt{N/2}} - z_{\alpha/4} \right) \right] \\ &= 1 - \left[ \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_i|}{\sigma/\sqrt{N}} + z_{\alpha/4} \right) - \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_i|}{\sigma/\sqrt{N}} - z_{\alpha/4} \right) \right] \end{aligned}$$

where the last equality stems from Assumption G.3. Since we are rejecting the null hypothesis of  $H_0 : \delta_1 = 0$  and  $\delta_0 = 0$  when the test in either subgroup shows significance, and the two tests are independent, the power of **GATE** is then

$$\begin{aligned} &1 - (1 - \xi_1)(1 - \xi_2) \\ &= 1 - \left[ \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_1|}{\sigma/\sqrt{N}} + z_{\alpha/4} \right) - \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_1|}{\sigma/\sqrt{N}} - z_{\alpha/4} \right) \right] \left[ \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_2|}{\sigma/\sqrt{N}} + z_{\alpha/4} \right) - \Phi \left( \frac{1}{\sqrt{2}} \frac{|\delta_2|}{\sigma/\sqrt{N}} - z_{\alpha/4} \right) \right] \end{aligned}$$

□

We now investigate three scenarios regarding the pattern of bias for the GATE estimators from the observational study:

**Scenario 1:** Only the GATE estimator for subgroup 1 is biased

This scenario can be depicted by letting  $\delta_1 = \delta \neq 0$  and  $\delta_2 = 0$ , so that we have  $\frac{\delta_1 + \delta_2}{2} = \frac{\delta}{2}$ . The power of **ATE** and **GATE** in this scenario can be given by, based on Propositions G.1 and G.2:

$$\xi_{\text{ATE}} = 1 - \left[ \Phi\left(\frac{|\delta/2|}{\sigma/\sqrt{N}} + z_{\alpha/2}\right) - \Phi\left(\frac{|\delta/2|}{\sigma/\sqrt{N}} - z_{\alpha/2}\right) \right] \quad (40)$$

$$\begin{aligned} \xi_{\text{GATE}} &= 1 - [\Phi(z_{\alpha/4}) - \Phi(-z_{\alpha/4})] \left[ \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right] \\ &= 1 - \left(1 - \frac{\alpha}{2}\right) \left[ \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right] \\ &= 1 - \left(1 - \frac{\alpha}{2}\right) \left[ \Phi\left(\sqrt{2} \frac{|\delta/2|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\sqrt{2} \frac{|\delta/2|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right] \end{aligned} \quad (41)$$

Denoting  $\delta^* := \frac{|\delta/2|}{\sigma/\sqrt{N}} \geq 0$ , we may simplify the expressions as,

$$\xi_{\text{ATE}} = 1 - [\Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2})] \quad (42)$$

$$\xi_{\text{GATE}} = 1 - \left(1 - \frac{\alpha}{2}\right) [\Phi(\sqrt{2}\delta^* + z_{\alpha/4}) - \Phi(\sqrt{2}\delta^* - z_{\alpha/4})] \quad (43)$$

Before we derive sufficient conditions for  $\xi_{\text{GATE}} > \xi_{\text{ATE}}$ , we state the following lemma on the properties of  $\Phi(\cdot)$  and  $\Phi^{-1}(\cdot)$ :

**Lemma G.1.**  $\forall \alpha \in (0, 1)$ ,  $\frac{z_{\alpha/4}}{z_{\alpha/2}} < \frac{z_{1/4}}{z_{1/2}} \approx 1.1185$ .

**Lemma G.2.**  $\forall a > 1$ ,  $\Phi(ax) - \Phi(x)$  is a strictly decreasing function in  $x$  as  $x > \sqrt{\frac{2 \log a}{a^2 - 1}}$ .

*Proof.* Taking the derivative of  $\Phi(ax) - \Phi(x)$  with respect to  $x$ , we have

$$\frac{\partial}{\partial x} [\Phi(ax) - \Phi(x)] = a\phi(ax) - \phi(x) = a \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2 x^2}{2}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [ae^{-\frac{a^2-1}{2}x^2} - 1]$$

When  $x > \sqrt{\frac{2 \log a}{a^2 - 1}}$ , we have, since  $a > 1$ ,

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [ae^{-\frac{a^2-1}{2}x^2} - 1] < \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [ae^{-\frac{a^2-1}{2}(\frac{2 \log a}{a^2-1})} - 1] = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [a \cdot \frac{1}{a} - 1] = 0$$

Therefore, at  $x > \sqrt{\frac{2 \log a}{a^2 - 1}}$ ,  $\Phi(ax) - \Phi(x)$  has strictly negative derivatives which implies it is strictly decreasing.  $\square$

Now we may derive the sufficient condition for  $\xi_{\text{GATE}} > \xi_{\text{ATE}}$ ,

$$\xi_{\text{GATE}} > \xi_{\text{ATE}}$$

$$\begin{aligned} \Leftrightarrow & \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) > \left(1 - \frac{\alpha}{2}\right) [\Phi(\sqrt{2}\delta^* + z_{\alpha/4}) - \Phi(\sqrt{2}\delta^* - z_{\alpha/4})] \\ \Leftarrow & \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi(\sqrt{2}\delta^* + z_{\alpha/4}) - \Phi(\sqrt{2}\delta^* - z_{\alpha/4}) \\ \Leftrightarrow & \Phi(\sqrt{2}\delta^* - z_{\alpha/4}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi(\sqrt{2}\delta^* + z_{\alpha/4}) - \Phi(\delta^* + z_{\alpha/2}) \\ \Leftarrow & \Phi(\sqrt{2}\delta^* - \sqrt{2}z_{\alpha/4}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi(\sqrt{2}\delta^* + \sqrt{2}z_{\alpha/4}) - \Phi(\delta^* + z_{\alpha/2}) \quad (\text{Lemma G.1}) \\ \Leftrightarrow & \Phi[\sqrt{2}(\delta^* - z_{\alpha/4})] - \Phi[\delta^* - z_{\alpha/2}] > \Phi[\sqrt{2}(\delta^* + z_{\alpha/4})] - \Phi[\delta^* + z_{\alpha/2}] \end{aligned}$$

Since  $\delta^* - z_{\alpha/2} < \delta^* + z_{\alpha/2}$ , from Lemma G.2, the last inequality holds as long as  $\delta^* - z_{\alpha/2} > \sqrt{\frac{2 \log(\sqrt{2})}{(\sqrt{2})^2 - 1}} = \sqrt{\log 2}$ .

That is, a sufficient condition for  $\xi_{\text{GATE}} > \xi_{\text{ATE}}$  is

$$\delta^* > \sqrt{\log 2} + z_{\alpha/2}$$

or, equivalently,

$$|\delta| > \frac{2\sigma}{\sqrt{N}}(\sqrt{\log 2} + z_{\alpha/2}) \quad (44)$$

Intuitively, we see from the above condition that as the magnitude of the bias in subgroup 1 increases or the sample size  $N$  increases, **GATE** will eventually have greater power than **ATE**.

**Scenario 2:** The **GATE** estimators for both subgroups are biased by the same magnitude but opposite direction

This scenario can be depicted by letting  $\delta_1 = \delta$  and  $\delta_2 = -\delta$ ,  $\delta \neq 0$ , so that we have  $\frac{\delta_1 + \delta_2}{2} = 0$ . Under which the power of **ATE** and **GATE** can be given by, based on Propositions G.1 and G.2:

$$\xi_{\text{ATE}} = 1 - [\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})] = \alpha \quad (45)$$

$$\xi_{\text{GATE}} = 1 - \left[ \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right]^2 \quad (46)$$

We may give a lower bound for  $\xi_{\text{GATE}}$ :

$$\xi_{\text{GATE}} = 1 - \left[ \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right]^2 > 1 - \left[ 1 - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right]^2 \quad (47)$$

Therefore, a sufficient condition for  $\xi_{\text{GATE}} > \xi_{\text{ATE}}$ , i.e. the power of **GATE** to be greater than **ATE** is,

$$|\delta| > \frac{\sigma}{\sqrt{N/2}}(z_{\alpha/4} + \Phi^{-1}(1 - \sqrt{1 - \alpha})) \quad (48)$$

which can be attained with a large enough bias magnitude  $|\delta|$  or large enough sample size  $N$  that overcomes the penalty of multiple testing.

**Scenario 3:** The **GATE** estimators for both subgroups are biased by the same magnitude and direction

This scenario can be depicted by letting  $\delta_1 = \delta_2 = \delta \neq 0$ , so that we have  $\frac{\delta_1 + \delta_2}{2} = \delta$ . Under which the power of **ATE** and **GATE** can be given by, based on Propositions G.1 and G.2:

$$\xi_{\text{ATE}} = 1 - \left[ \Phi\left(\frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/2}\right) - \Phi\left(\frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/2}\right) \right] \quad (49)$$

$$\xi_{\text{GATE}} = 1 - \left[ \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}} \frac{|\delta|}{\sigma/\sqrt{N}} - z_{\alpha/4}\right) \right]^2 \quad (50)$$

Denoting  $\delta^* := \frac{|\delta|}{\sigma/\sqrt{N}} \geq 0$ , we may simplify the expressions as,

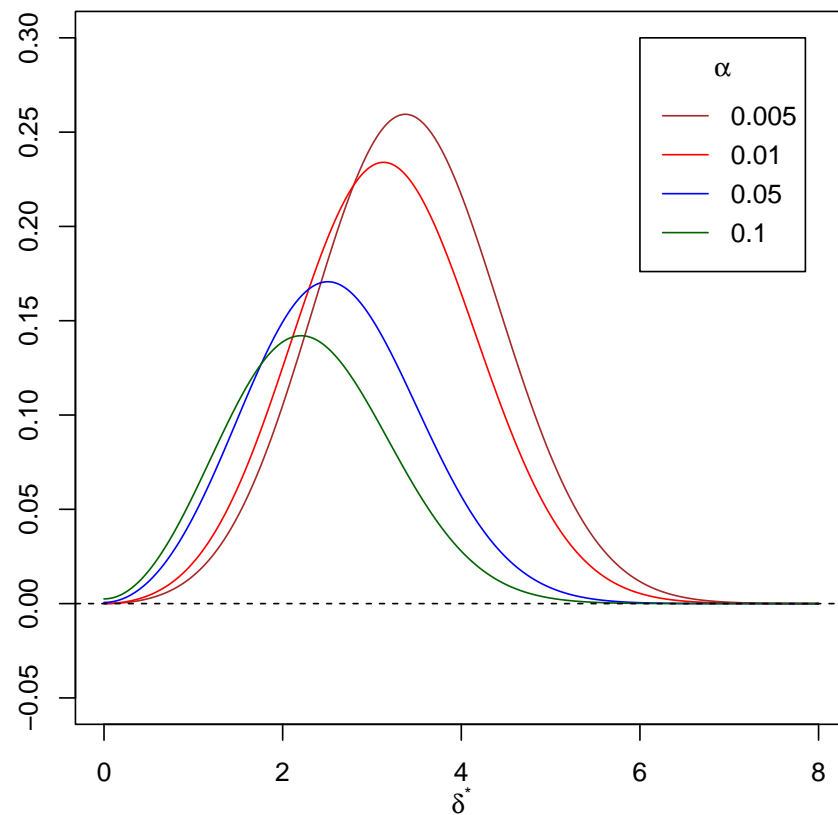
$$\xi_{\text{ATE}} = 1 - \left[ \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) \right] \quad (51)$$

$$\xi_{\text{GATE}} = 1 - \left[ \Phi\left(\frac{1}{\sqrt{2}}\delta^* + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\delta^* - z_{\alpha/4}\right) \right]^2 \quad (52)$$

Therefore, the condition for  $\xi_{\text{ATE}} > \xi_{\text{GATE}}$  is equivalent to

$$g(\delta^*) := \left[ \Phi\left(\frac{1}{\sqrt{2}}\delta^* + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\delta^* - z_{\alpha/4}\right) \right]^2 - \left[ \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) \right] > 0 \quad (53)$$





**Figure 5:** Plot of function  $g(\cdot)$  under  $\alpha = 0.005, 0.01, 0.05$  and  $0.1$

A graph for  $g(\delta^*)$  with  $\alpha = 0.005, 0.01, 0.05, 0.1$  is shown in Figure 5, which demonstrates that  $g(\delta^*) > 0$  is satisfied for any  $\delta^* > 0$ . Therefore, under the scenario where a common bias is shared across subgroups, the power of **ATE** is greater than **GATE** irrespective of the magnitude of bias.

Overall, we find that the relative asymptotic power of **ATE** and **GATE** depends on the homogeneity of bias amongst the subgroups and the magnitude of the bias, and should be analyzed on a case-by-case basis.