Hidden among subgroups: Detecting critical treatment effect bias in observational studies

Piersilvio De Bartolomeis*, Javier Abad*, Konstantin Donhauser*, and Fanny Yang

Department of Computer Science, ETH Zürich

Abstract

Randomized trials are considered the gold standard for making informed decisions in medicine, yet they often lack generalizability to the patient populations in clinical practice. Observational studies, on the other hand, cover a broader patient population but are prone to various biases. Thus, before using an observational study for decision-making, it is crucial to *benchmark* its treatment effect estimates against those derived from a randomized trial. We propose a novel strategy to benchmark observational studies beyond the average treatment effect. First, we design a statistical test for the null hypothesis that the treatment effects, conditioned on a subset of relevant features, differ up to some tolerance value. Then, we estimate an asymptotically valid lower bound on the maximum bias strength for any subgroup in the observational study. We validate our lower bound in a real-world setting and show that it leads to conclusions that align with established medical knowledge[1].

1 Introduction

Randomized trials have traditionally been the gold standard for informed decision-making in medicine, as they allow for unbiased estimates of treatment effects under mild assumptions. However, there is often a significant discrepancy between the patients observed in clinical practice and those enrolled in randomized trials [11]. These distribution shifts compromise the generalizability of the trials to broader populations [44].

The U.S. Food and Drug Administration currently promotes using observational data when randomized data provides limited evidence, as it is usually more representative of the patient population in clinical practice [31, 40]. Nonetheless, several sources of bias, such as unobserved confounding, can significantly compromise the causal conclusions drawn from non-randomized data. Hence, it is crucial to assess the quality of observational data before using it for any downstream medical task.

Benchmarking observational studies has become a popular strategy to assess the reliability of observational data when a randomized trial is available [5, 12]. The main idea behind this approach is to first emulate the procedures adopted in the randomized trial within the observational study, for example, by using the framework in Hernán and Robins [19]. Then, the treatment effect estimates from the emulated observational study are compared with those from the randomized trial. If the estimates are similar, we may be willing to trust the observational study results for patient populations where the randomized data is insufficient.

*These authors contributed equally.

[1]See our GitHub repository for the source code: `https://github.com/jaabmar/kernel-test-bias`.

To support the benchmarking framework, several works propose statistical tests that compare treatment effect estimates between randomized and observational data [8, 10, 24, 53, 56]. In particular, two properties have been identified in the literature as essential for effective benchmarking of observational studies: *granularity* and *tolerance*. Granularity allows the detection of bias on a subgroup or individual level, thereby improving the power of benchmarking. Tolerance allows the acceptance of studies with negligible bias that does not impact decision-making, thereby reducing false rejections. However, to date, no existing statistical test satisfies both properties.

Contributions We design a statistical test for the null hypothesis that treatment effects differ up to some tolerance value when conditioned on a relevant subset of features. Our test is the first, to our knowledge, to satisfy granularity and offer tolerance. Further, we use our test to estimate an asymptotically valid lower bound on the maximum bias strength for any subgroup in the observational study. We validate our lower bound using real-world data and show that it leads to conclusions that align with current medical knowledge.

2 Problem setting

We have access to two datasets, D_{rct} of size $2n_{\text{rct}}$ from a randomized trial (rct) and D_{os} of size $2n_{\text{os}}$ from an observational study (os), containing tuples $Z := (X, Y, T)$ of observed covariates $X \in \mathbb{R}^d$, observed outcomes $Y \in \mathbb{R}$ and treatment assignment variable $T \in \{0, 1\}$. We assume that the data is drawn i.i.d from the distributions $\mathbb{P}^{\text{rct}}$ and $\mathbb{P}^{\text{os}}$. Additionally, we assume that $\mathbb{P}^{\diamond}$ is the marginal distribution of the full distribution $\mathbb{P}^{\diamond}_{\text{full}}$ over $(X, U, Y(0), Y(1), Y, T)$ for $\diamond \in \{\text{rct}, \text{os}\}$, where $U \in \mathbb{R}^k$ is a vector of unobserved confounders and $(Y(0), Y(1))$ are real-valued potential outcomes.

Our goal is to detect bias in the treatment effect estimates derived from the observational study. To this end, the randomized trial must provide valid estimates, as captured by the following standard assumption.

Assumption 2.1. *The data-generating process of the randomized trial is internally valid, i.e.* $\mathbb{P}^{\text{rct}}_{\text{full}}$ *satisfies*

$$\begin{aligned} &(i) \;\; Y = Y(T) \;\; \mathbb{P}^{\text{rct}}_{\text{full}} - \text{almost surely}, \\ &(ii) \;\; T \perp\!\!\!\perp (Y(1), Y(0)), \\ &(iii) \;\; \mathbb{P}^{\text{rct}}_{\text{full}}(T = 1 \mid X, U) = \pi \in (0, 1). \end{aligned}$$

These assumptions are widely adopted in causal inference and are expected to hold by design in a completely randomized experiment [45]. Nevertheless, we remark that the results in this paper also apply if we replace the randomized trial with an observational study that is unaffected by unobserved confounding.

2.1 Null hypothesis

Heterogeneous effect estimation plays a crucial role in medicine, as it can improve the understanding of variations in the patient population. The most common target quantity in this context is the conditional average treatment effect (CATE), given by

$$\mu^{\text{rct}}(X) := \mathbb{E}_{\mathbb{P}^{\text{rct}}_{\text{full}}}\{Y(1) - Y(0) \mid X\} \quad \text{and} \quad \mu^{\text{os}}(X) := \mathbb{E}_{\mathbb{P}^{\text{os}}_{\text{full}}}\{Y(1) - Y(0) \mid X\}.$$

The CATE is typically estimated by the difference between two regression functions that we denote by

$$\tau^{\diamond}(X) = \mathbb{E}_{\mathbb{P}^{\diamond}}\{Y \mid T = 1, X\} - \mathbb{E}_{\mathbb{P}^{\diamond}}\{Y \mid T = 0, X\}, \quad \text{for } \diamond \in \{\text{rct}, \text{os}\}.$$

However, while the CATE in the randomized trial is identifiable under Assumption 2.1, i.e. $\tau^{\text{rct}}(X) = \mu^{\text{rct}}(X)$ point-wise, the same is not true for the CATE in the observational study, that is generally not identifiable.

Thus, we would like to test if the treatment effects estimated in the observational study significantly differ from those in the randomized trial. More specifically, we would like to compare treatment effects conditioned on relevant features since researchers are often interested in heterogeneity w.r.t. a small subset of features rather than the full set required for confounding adjustment (see e.g. Abrevaya et al. [1]). Formally, this setting is captured by the null hypothesis

$$H_0: \ \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\tau^{\text{rct}}(X) \mid X^{\mathcal{J}}\right\} \in \left[\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\tau_{-}^{\text{os}}(X) \mid X^{\mathcal{J}}\right\}, \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\tau_{+}^{\text{os}}(X) \mid X^{\mathcal{J}}\right\}\right], \ \mathbb{P}^{\text{rct}}_{X^{\mathcal{J}}} - \text{almost surely}, \tag{1}$$

where $X^{\mathcal{J}}$ is a subset of covariates with indices $\mathcal{J} \subseteq \{1, \cdots, d\}$, and $\tau_{-}^{\text{os}} \leq \tau^{\text{os}} \leq \tau_{+}^{\text{os}}$ holds point-wise. The tolerance functions $\tau_{\pm}^{\text{os}}$ capture how much the treatment effects can differ. Below, we discuss two concrete examples.

Example 1: User-specified tolerance A natural choice for the tolerance functions is to simply add (respectively subtract) a user-specified tolerance function $\delta(X) \geq 0$,

$$\tau_{\pm}^{\text{os}}(X) = \tau^{\text{os}}(X) \pm \delta(X).$$

The tolerance function $\delta(X)$ can incorporate all sources of bias in the observational study, such as unobserved confounding and non-adherence to treatment assignments. Similar tolerance functions have been previously used in the context of modeling violations of the transportability assumptions, see, e.g. Dahabreh et al. [6, 7], Nguyen et al. [37, 38].

Example 2: Sensitivity analysis bounds Another practical choice for the tolerance functions $\tau_{\pm}^{\text{os}}$ is to use the upper and lower bounds arising from a sensitivity analysis model. For instance, the marginal sensitivity model introduced by Tan [50] is commonly used to account for unobserved confounding in observational data. In particular, this model assumes that U has a limited influence on T, that is

$$\Gamma^{-1} \leq \frac{\mathbb{P}^{\text{os}}_{\text{full}}(T=1|X,U)}{\mathbb{P}^{\text{os}}_{\text{full}}(T=0|X,U)} \Big/ \frac{\mathbb{P}^{\text{os}}(T=1|X)}{\mathbb{P}^{\text{os}}(T=0|X)} \leq \Gamma, \quad \mathbb{P}^{\text{os}}_{\text{full}} - \text{almost surely}.$$

Under the marginal sensitivity model, we can estimate an interval for the treatment effect that depends on an assumed *confounding strength* Γ. We can then define $\tau_{\pm}^{\text{os}}$ as the upper and lower bounds for τ^{os} under the assumption of Γ-bounded confounding strength. Several estimators have recently emerged in the literature for sensitivity analysis bounds [26, 28, 39]. Further, if we are willing to assume that transportability holds, i.e. $\mu^{\text{rct}}(X) = \mu^{\text{os}}(X)$ point-wise, we can test whether the marginal sensitivity model is well-specified [8].

2.2 Related work

In this section, we discuss related works that combine randomized and non-randomized data either to detect flawed observational studies or to obtain more reliable treatment effect estimates.

Statistical tests based on average treatment effects Given the challenges associated with estimating treatment effects with non-randomized data, several works propose to detect bias in observational studies using randomized trials [14, 34, 53, 56]. In particular, they introduce statistical tests for the null hypothesis

$$H_0 : \mathbb{E}_{\mathbb{P}^{\text{rct}}_X}\{\tau^{\text{rct}}(X)\} = \mathbb{E}_{\mathbb{P}^{\text{rct}}_X}\{\tau^{\text{os}}(X)\}, \tag{2}$$

where $\mathbb{P}_X^{\text{rct}}$ is the marginal distribution of X under $\mathbb{P}^{\text{rct}}$. In our setting, rejecting H_0 implies that either the treatment effect estimate from the observational study is biased or the transportability assumption is violated, i.e. $\mu^{\text{rct}}(X) \neq \mu^{\text{os}}(X)$ for some X. However, a major limitation of testing the null hypothesis in Equation (2) is that, even in infinite samples, we reject observational studies with negligible treatment effect bias. This approach can be too restrictive in real-world settings, where some bias is likely present.

Statistical tests with tolerance One way to address the restrictiveness of previous statistical tests and reduce false rejections is to incorporate some tolerance. More formally, given some user-specified tolerance functions $\tau_{\pm}^{\text{os}}$, De Bartolomeis et al. [8] propose a test for the null hypothesis

$$H_0 : \mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\{\tau^{\text{rct}}(X)\} \in \left[\mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\{\tau_{-}^{\text{os}}(X)\}, \mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\{\tau_{+}^{\text{os}}(X)\}\right], \tag{3}$$

where $\tau_{-}^{\text{os}} \leq \tau^{\text{os}} \leq \tau_{+}^{\text{os}}$ holds point-wise. For instance, if we choose sensitivity analysis bounds as tolerance functions and assume transportability ($\mu^{\text{rct}}(x)=\mu^{\text{os}}(x)$ for all $x \in \mathbb{R}^d$), we can test for the presence of unobserved confounding above a certain strength. However, a limitation of statistical tests based on the null hypotheses in Equations (2) and (3) is that they are not granular, i.e. they cannot detect bias on a subgroup or individual level. In particular, bias may cancel out on average, leading to flawed studies being accepted. In contrast, our null hypothesis in Equation (1) also satisfies granularity and is more general, i.e. it recovers existing tests with tolerance when the subset of features J is the empty set.

Statistical tests with granularity Several works have addressed the lack of granularity in statistical tests based on average treatment effects. Hussain et al. [25] compare group-level treatment effects using pre-specified subgroups; however, this approach suffers from multiple testing issues and cannot detect bias at the individual level. More recently, Hussain et al. [24] propose a kernel test for the null hypothesis

$$H_0 : \tau^{\text{rct}}(X) = \tau^{\text{os}}(X) \quad \mathbb{P}_X^{\text{rct}} - \text{almost surely.} \tag{4}$$

The main advantage of such a test is that it can detect bias in arbitrarily fine-grained subpopulations. Further, Demirel et al. [10] extend the above test to account for right-censored outcomes. However, all the statistical tests with granularity fall short of incorporating tolerance functions. In contrast, our null hypothesis in Equation (1) offers both tolerance and granularity by choosing a relevant subset of features J. Notably, when the tolerance functions are the same ($\tau_{-}^{\text{os}}(x) = \tau_{+}^{\text{os}}(x)$ for all $x \in \mathbb{R}^d$), and we consider the entire feature vector ($X^{\mathcal{J}} = X$), our null hypothesis recovers the one in Equation (4).

Combining data for estimation A recent line of work proposes to combine randomized and observational data to estimate treatment effects [4, 27, 43, 54, 55, 56]. In particular, the work of Cheng and Cai [3] is closely related to ours, where the authors use kernel regression to estimate the treatment effect conditional on a subset of features. This line of research focuses on learning and correcting the bias between observational and randomized estimates. Such approaches are particularly valuable when the support of the two studies is the same, as they can reduce the variance of the treatment effect estimates by pooling the data. However, when the supports are different, learning the bias requires strong parametric assumptions for extrapolation. In contrast, the goal of statistical tests is to identify flawed observational studies; see, e.g. Forbes and Dahabreh [12]. This task is feasible even in settings where the supports do not match, as it is enough to detect differences in the common support of the two studies.

3 Methodology

We now present our methodology for testing the null hypothesis in Equation (1). More concretely, we propose testing for a slightly more restrictive null hypothesis that reduces the problem to testing equality of conditional expectations, i.e.

$$H_0^{\mathcal{G}}: \quad \exists g^\star \in \mathcal{G} \quad \text{s.t.} \quad \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\tau^{\text{rct}}(X) \mid X^{\mathcal{J}}\right\} = \mathbb{E}_{\mathbb{P}^{\text{rct}}}\{\tau^{\text{os}}_{g^\star}(X) \mid X^{\mathcal{J}}\}, \quad \mathbb{P}^{\text{rct}}_{X^{\mathcal{J}}} - \text{almost surely}, \tag{5}$$

where $\mathcal{G}$ is a class of functions from $\mathbb{R}^{|J|} \to [0,1]$, and for any $g \in \mathcal{G}$, we define $\tau^{\text{os}}_g(X) := g\left(X^{\mathcal{J}}\right)\tau^{\text{os}}_+(X) + \left(1 - g\left(X^{\mathcal{J}}\right)\right)\tau^{\text{os}}_-(X)$. Observe that by testing $H_0^{\mathcal{G}}$ instead of H_0 in Equation (1), we additionally assume that $\mathcal{G}$ is a sufficiently rich function class to accurately model the dependence of the bias (captured via $g^\star$) on $X^{\mathcal{J}}$. In practice, one can either restrict $\mathcal{G}$ to a particular function class if domain knowledge is available or use neural networks as general function approximators (for which the assumption is expected to hold).

We remark that while the problem of testing equality of conditional expectations has been extensively studied [9, 33, 35, 36, 42], to our knowledge, incorporating a tolerance into such tests has not been previously explored. In what follows, we first propose an oracle test statistic that assumes the tolerance functions $\tau^{\text{os}}_\pm$ are known. Then, we provide asymptotic guarantees for the finite-sample test statistic where the tolerance functions are estimated.

3.1 Oracle test statistic

We define a *signal* function that captures the bias between the two data sources. More formally, we recall that $Z = (X, Y, T)$ and define

$$\psi_g(Z) = Y\left(\frac{T}{\pi} - \frac{1-T}{1-\pi}\right) - \tau^{\text{os}}_g(X). \tag{6}$$

Then, we observe that the conditional expectation

$$\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\tau^{\text{rct}}(X) \mid X^{\mathcal{J}}\right\} = \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{Y\left(\frac{T}{\pi} - \frac{1-T}{1-\pi}\right) \mid X^{\mathcal{J}}\right\},$$

where the equality follows from the classic result in Horvitz and Thompson [22]. Thus, we can rewrite the null hypothesis in Equation (5) as

$$H_0^{\mathcal{G}}: \exists g^\star \in \mathcal{G} \quad \text{s.t.} \quad \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[\psi_{g^\star} \mid X^{\mathcal{J}}\right] = 0, \quad \mathbb{P}^{\text{rct}}_{X^{\mathcal{J}}} - \text{almost surely}.$$

If $g^\star$ were known, we could simply use any (kernel) conditional moment test. Indeed, observe that the null hypothesis $H_0^{\mathcal{G}}$ implies an infinite set of unconditional moment restrictions, i.e. for any $g \in \mathcal{G}$ it holds that

$$\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[\psi_g(Z) \mid X^{\mathcal{J}}\right] = 0, \;\; \mathbb{P}^{\text{rct}}_{X^{\mathcal{J}}} - \text{almost surely} \implies \mathbb{E}_{\mathbb{P}^{\text{rct}}}[\psi_g(Z) f(X^{\mathcal{J}})] = 0,$$

for all measurable functions f. Since testing the implication above for all measurable functions is infeasible, Muandet et al. [35] propose restricting f to be in a reproducing kernel Hilbert space (RKHS). The problem then becomes more tractable since it holds that

$$\begin{aligned}\mathbb{H}^2(\psi_g) := \left(\sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\psi_g(Z) f(X^{\mathcal{J}})\right\}\right)^2 &= \left\|\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\psi_g(Z) k(X^{\mathcal{J}}, \cdot)\right\}\right\|^2_{\mathcal{F}} \\ &= \mathbb{E}_{\mathbb{P}^{\text{rct}}}\{\psi_g(Z) k(X^{\mathcal{J}}, X_2^{\mathcal{J}}) \psi_g(Z_2)\},\end{aligned} \tag{7}$$

where k is the reproducing kernel corresponding to an RKHS $\mathcal{F}$, and $X_2^{\mathcal{J}}, Z_2$ are independent copies of $X^{\mathcal{J}}, Z$ with the same distribution. Therefore, under the null hypothesis $H_0^{\mathcal{G}}$, it holds that $\mathbb{H}^2(\psi_g) = 0$ if $g = g^\star$. Further, an unbiased empirical estimate of $\mathbb{H}^2(\psi)$ is the classical U-statistic [47]

$$\frac{1}{n(n-1)} \sum_{i \neq j} \psi(Z_i) k(X_i^{\mathcal{J}}, X_j^{\mathcal{J}}) \psi(Z_j).$$

Under the null hypothesis that $\mathbb{H}^2(\psi) = 0$, the U-statistic converges in distribution to a weighted χ^2-statistic, and its quantile can be consistently estimated via bootstrapping [23]. However, in our setting, bootstrapping the test statistic requires knowing $\psi_{g^\star}$, and thus, this approach is only feasible when $g^\star$ is known.

When $g^\star$ is unknown, we can instead use the cross U-statistic from Kim and Ramdas [30]. Formally, we split the dataset D_{rct} equally into two folds of size n_{rct}; the test statistic is then given by

$$\hat{\mathbb{H}}^2(\psi) = \frac{1}{n_{\text{rct}}} \sum_{i=1}^{n_{\text{rct}}} f(Z_i; \psi), \tag{8}$$

where $f(Z_i; \psi) := \frac{1}{n_{\text{rct}}} \sum_{k=n_{\text{rct}}+1}^{2n_{\text{rct}}} \psi(Z_i) k(X_i^{\mathcal{J}}, X_k^{\mathcal{J}}) \psi(Z_k)$. The main advantage of the cross U-statistic is that it is asymptotically normal under the null hypothesis in Equation (5) and weak regularity assumptions on ψ and k (see Theorem 3.1), i.e.

$$\sqrt{n_{\text{rct}}}\, \hat{\mathbb{H}}^2(\psi_{g^\star}) \to \mathcal{N}\left(0, \sigma^2\left(\psi_{g^\star}\right)\right),$$

where the variance term is given, for any $g \in \mathcal{G}$, by

$$\sigma^2\left(\psi_g\right) = \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\left(f(Z; \psi_g) - \mathbb{E}_{\mathbb{P}^{\text{rct}}}\{f(Z; \psi_g)\}\right)^2\right\}.$$

While the variance still depends on $g^\star$, we can obtain a valid test by solving the optimization problem

$$\min_{g \in \mathcal{G}} \left| \frac{\sqrt{n_{\text{rct}}}\, \hat{\mathbb{H}}^2(\psi_g)}{\hat{\sigma}\left(\psi_g\right)} \right|, \tag{9}$$

where $\hat{\sigma}\left(\psi_g\right)$ is the finite sample estimate of the variance term $\sigma^2\left(\psi_g\right)$. In the next section, we formally show that the test statistic in Equation (9) is asymptotically valid, even when the tolerance functions $\tau_{\pm}^{\text{os}}$ are estimated from the data.

3.2 Theoretical guarantees

Since, in practice, we do not have access to the signal function ψ_g, we define the finite-sample analogous as

$$\hat{\psi}_g(Z) = Y\left(\frac{T}{\pi} - \frac{1-T}{1-\pi}\right) - \hat{\tau}_g^{\text{os}}(X),$$

where $\hat{\tau}^{\text{os}}(X) = g(X^{\mathcal{J}})\hat{\tau}_+^{\text{os}}(X) + \left(1 - g\left(X^{\mathcal{J}}\right)\right)\hat{\tau}_-^{\text{os}}(X)$ is estimated using only the observational data D_{os}. We can then define the testing function $\hat{\phi}$ to be

$$\hat{\phi}(\alpha) := \mathbb{I}\left\{ \min_{g \in \mathcal{G}} \left| \frac{\sqrt{n_{\text{rct}}}\, \hat{\mathbb{H}}^2(\hat{\psi}_g)}{\hat{\sigma}\left(\hat{\psi}_g\right)} \right| \geq z_{1-\alpha} \right\}, \tag{10}$$

where z_α is the α-quantile of the half-normal distribution.

We now provide sufficient conditions for $\hat{\phi}$ to be an asymptotically valid test.

Theorem 3.1 (Validity of the test)**.** *We make the following assumptions:*

(i) the upper and lower bounds $\tau_{\pm}^{\text{os}}$ are uniformly bounded and for some constant $M_Y > 0$, the observed outcome has bounded conditional moments

$$\mathbb{E}_{\mathbb{P}^{\text{rct}}}[Y^4|X] \leq M_Y, \quad \mathbb{P}_X^{\text{rct}} - \text{almost surely.} \tag{11}$$

(ii) the kernel k is uniformly upper bounded by a constant $M_k > 0$.

(iii) the variance term is non-zero, i.e.

$$\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{ \psi_{g^\star}^2(Z)\ k^2(X^{\mathcal{J}}, X_2^{\mathcal{J}})\ \psi_{g^\star}^2(Z_2)\right\} > 0, \tag{12}$$

where $X_2^{\mathcal{J}}$ and Z_2 are independent copies of $X^{\mathcal{J}}$ and Z with the same distribution.

(iv) the estimates $\hat{\tau}_{\pm}^{\text{os}}$ are trained only on the observational data D_{os} and satisfy

$$\|\tau_{\pm}^{\text{os}} - \hat{\tau}_{\pm}^{\text{os}}\|_{L_2(\mathbb{P}^{\text{rct}})} = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n_{\text{rct}}}}\right). \tag{13}$$

Then, under the setting described in Section 2, the testing function $\hat{\phi}(\alpha)$ from Equation (10) is a valid asymptotic test at level α for the null hypothesis $H_0^{\mathcal{G}}$ from Equation (5).

We refer the reader to Appendix A.2 for a complete proof.

Discussion of assumptions Assumptions (*i-iii*) are mild and apply to very general settings. For instance, they are satisfied when Y is a bounded, non-deterministic random variable, i.e. the variance is non-zero, and k is a standard radial basis function kernel such as the Laplacian or Gaussian kernels. Assumption (*iv*) is stronger and generally only expected to hold when $n_{\text{os}} \gg n_{\text{rct}}$ and the support of the randomized control trial is contained in the support of the observational study, i.e. $\text{supp}(\mathbb{P}_X^{\text{rct}}) \subseteq \text{supp}(\mathbb{P}_X^{\text{os}})$. These two conditions are realistic in our setting, as they align with the standard design of observational studies [13, 17, 46]. Nevertheless, we remark that previous works either assume oracle access to the tolerance functions $\tau_{\pm}^{\text{os}}$ [10, 24] or impose similar assumptions on the estimation rates [8].

Power of the test While Theorem 3.1 only shows asymptotic validity, we further present guarantees for the asymptotic power of the test in Appendix A.1. In particular, in Theorem A.1, we show that under the alternative hypothesis

$$H_A^{\mathcal{G}} : \inf_{g \in \mathcal{G}} \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left\{\psi_g(Z) f(X^{\mathcal{J}})\right\} > 0,$$

the test statistic in Equation (10) grows at the typical rate of order $\sqrt{n_{\text{rct}}}$ for a fixed function class $\mathcal{G}$. Thus, it yields the same asymptotic power (up to constant factors) as a standard conditional moment test without tolerance, i.e. where $\tau_{-}^{\text{os}} = \tau_{+}^{\text{os}}$ holds point-wise; see, e.g. Hussain et al. [24], Muandet et al. [35].

Asymptotically valid lower bound on the bias For the sake of clarity, let us consider the tolerance functions from Example 1, i.e. $\tau_{\pm}^{\text{os}}(X) = \tau^{\text{os}}(X) \pm \delta$, for some constant $\delta \in \mathbb{R}^+$. Given the result in Theorem 3.1, we can construct an asymptotically valid lower bound on the maximum amount of bias in the treatment effect estimated from the observational data.

More formally, we define a data-dependent lower bound on the bias using the testing function in Equation (10)

$$\hat{\delta}_{\text{LB}} := \inf_{\delta} \{\delta : \hat{\phi}(\alpha) = 0\}, \tag{14}$$

where $\hat{\phi}$ depends implicitly on δ via the tolerance functions. Further, we define the maximum point-wise bias in the observational data as

$$\delta^\star := \|\mathbb{E}_{\mathbb{P}^{\text{rct}}}\{\tau^{\text{rct}}(X) \mid X^{\mathcal{J}}\} - \mathbb{E}_{\mathbb{P}^{\text{rct}}}\{\tau^{\text{os}}(X) \mid X^{\mathcal{J}}\}\|_\infty.$$

Then, it holds that

$$\mathbb{P}^{\text{rct}}\left(\delta^\star \geq \hat{\delta}_{\text{LB}}\right) = 1 - \alpha + o(1).$$

A similar quantity was already proposed in De Bartolomeis et al. [8] to lower bound the unobserved confounding strength. In contrast, we focus on a lower bound for the bias, and we show in the following sections how it can be used for decision-making. Further, we remark that the lower bound defined in Equation (14) is optimistic, as the bias could be arbitrarily high outside the support of the randomized trial.

4 Semi-synthetic experiments

In this section, we evaluate our test and the resulting lower bound in finite-sample semi-synthetic experiments.

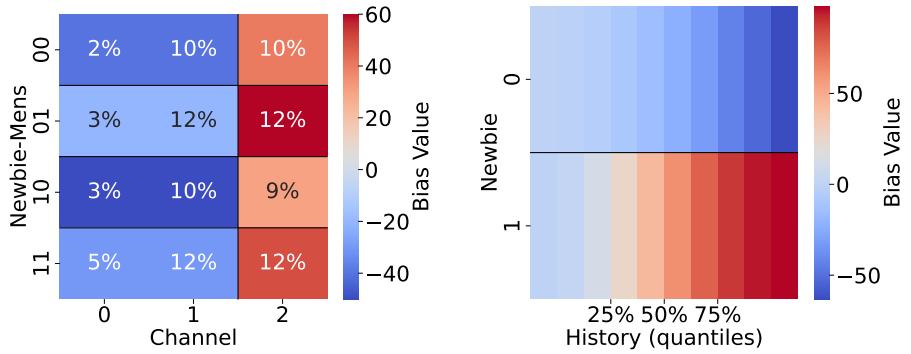
4.1 Experimental setting

Dataset We evaluate our testing procedure on a semi-synthetic dataset derived from a real-world randomized trial: Hillstrom's MineThatData Email dataset [20]. The Hillstrom dataset contains records of 64,000 customers who made purchases online within the last twelve months. We consider a combined treatment group, which constitutes approximately 66% of the dataset, and a control group. The outcome represents the dollars spent in the two weeks post-campaign. The dataset provides information on individual annual spending, newcomer status, and geographical location, among others. We normalized continuous features and one-hot-encoded categorical features, resulting in a 13-dimensional dataset. By default, we use 80% of the full dataset as the observational study (os) and the remaining 20% as the randomized trial (rct).

Bias model We consider two different models for the bias between studies, given by $\delta^\star(x) = \tau^{\text{rct}}(x) - \tau^{\text{os}}(x)$, for all $x \in \mathbb{R}^d$. We illustrate the two scenarios in Figure 1. In scenario 1 (Figure 1a), we add biases of varying magnitudes across 12 subgroups defined by combinations of the binary features `newbie` and `mens` and the categorical feature `channel`. The largest bias is $\delta^\star = 60$, and it affects only 12% of the observational dataset. The subgroup biases roughly cancel each other out on average, resulting in an average bias close to zero, i.e. $\mathbb{E}\{\delta^\star(X)\} \approx 0$. In scenario 2 (Figure 1b), we model the bias as a quadratic polynomial of the feature `history`, sampling different coefficients for the two values of `newbie` from a normal distribution.

User-defined tolerance and baselines We refer to the testing function proposed in this paper as $\hat{\phi}^{\text{CATE}}$, and we instantiate it using constant upper and lower bounds for the tolerance function, as described in Example 1 from Section 2.1 ($\tau^{\text{os}}_{\pm}(X) = \tau^{\text{os}}(X) \pm \delta$ for some constant $\delta \in \mathbb{R}^+$). We compare our test against $\hat{\phi}^{\text{ATE}}$, which is a slight modification[2] of the test with tolerance proposed in De Bartolomeis et al. [8]. For

[2] $\hat{\phi}^{\text{ATE}}$ is a t-test for the null hypothesis that average treatment effects between the studies differ at most δ.



(a) Scenario 1 (b) Scenario 2

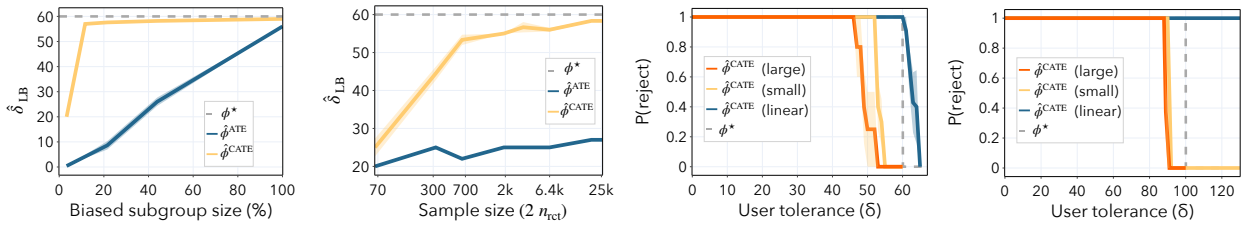
Figure 1: Heatmap visualizations of the bias for (a) Scenario 1 based on 12 subgroups with different biases (the numbers in the cells represent the percentage w.r.t. the full observational dataset), and (b) Scenario 2 based on a quadratic polynomial bias.

both tests, we can compute the lower bound on the bias $\hat{\delta}_{\text{LB}}$, as defined in Equation (14). Note that while our method allows us to select a subset of features $X^{\mathcal{J}}$ that are interesting for the heterogeneity of the treatment effect, we use the full feature set in all our semi-synthetic experiments. We thus show the effectiveness of our test even when considering a relatively large set of features, and we expect power to improve when considering a smaller subset; see e.g. the ablation studies for $X^{\mathcal{J}}$ in Appendix C.1.

Implementation To compute the test statistic, we use the Laplacian kernel with a scale of 1.0 in all experiments. We perform gradient descent for 6000 epochs using the `Adam` optimizer from the JAX-based library `optax` with its default hyperparameters and record the smallest test statistic. As function class $\mathcal{G}$, we consider linear functions and two multilayer perceptrons (MLPs), one *small* and one *large*, with hidden layer widths of 10 and 100-50-10-5 neurons, respectively. For the linear function and the small MLP, we set the learning rate to 0.1, and for the large MLP, we set it to 0.01. For the test $\hat{\phi}^{\text{ATE}}$, we use 500 bootstrap samples to estimate the variance of the test statistic.

4.2 Experimental results

We now discuss our experimental results, depicted in Figure 2. We first conduct ablation studies for a simplified setting of Scenario 1, where only one subgroup has a constant bias of $\delta^* = 60$, while the rest remain unbiased. We study the effect of the biased subgroup size (Figure 2a) and the randomized trial sample size (Figure 2b) on the lower bounds $\hat{\delta}_{\text{LB}}$ obtained from our test and the baseline $\hat{\phi}^{\text{ATE}}$. Next, we assess the validity and power of our test in two more complex scenarios (Figures 2c and 2d). An important consideration is the selection of the function class $\mathcal{G}$ in practice; it should be sufficiently large to encompass $g^\star$, but overly large function classes may result in a more difficult optimization problem. Thus, we also conduct ablation studies for $\mathcal{G}$. Our results show that granularity significantly improves the power of the test and, consequently, the estimated lower bound on the bias: $\hat{\phi}^{\text{CATE}}$ consistently outperforms the baseline across all scenarios and demonstrates robustness w.r.t. the choice of function class in the ablation studies.



(a) Increasing subgroup bias (b) Increasing n_{rct} (c) Scenario 1 (d) Scenario 2

Figure 2: For all the plots: the significance level is set at $\alpha = 0.05$, $\phi^\star$ denotes the oracle test, which rejects for $\delta < \delta^*$. (a-b) Simplified setting of Scenario 1 with a single subgroup having constant bias $\delta^* = 60$: we plot the bias lower bound $\hat{\delta}_{\text{LB}}$ as a function of (a) the biased subgroup percentage w.r.t. total sample size and (b) the randomized trial sample size. (c-d) Probability of rejection for different function classes $\mathcal{G}$ as a function of the user-specified tolerance δ for (c) scenario 1 (Figure 1a) based on 12 subgroups with different biases and (d) scenario 2 (Figure 1b) based on a quadratic polynomial bias. We report mean and standard error over 5 runs. The coefficients for the polynomial bias are fixed across runs.

Effect of biased subgroup and rct sample sizes Figure 2a shows that our test yields an average lower bound $\hat{\delta}_{\text{LB}}$ close to the true maximum bias δ^*. This implies that the test remains valid and exhibits significant power, even when the biased subgroup represents roughly 14% of the observational dataset. In contrast, $\hat{\phi}^{\text{ATE}}$ experiences a significant drop in power as the proportion of biased data points decreases. Such behavior is expected since $\hat{\phi}^{\text{ATE}}$ only tests for the difference of averages, and it cannot detect bias in small subgroups, i.e. it is not granular. In Figure 2b, we add a constant bias of 60 to 44% of the observational data points and study the effect of the randomized trial sample size. While our test suffers more than $\hat{\phi}^{\text{ATE}}$ from a decrease in the sample size due to the use of kernels, it always yields higher power, including in the very small sample size regime with only 70 data points. These results show the importance of granularity: even in simple settings, $\hat{\phi}^{\text{ATE}}$ can fail to flag significantly biased datasets, in contrast to our method.

Validity and power in complex scenarios Figure 2c and Figure 2d show the validity and power of our testing procedure for Scenario 1 (illustrated in Figure 1a) and Scenario 2 (illustrated in Figure 1b), respectively. In both scenarios, if we use a neural network to approximate the bias function, our test remains valid and shows very high power since it rejects the null hypothesis at values of δ close to the true bias $\delta^\star$.

Effect of misspecified function class Notably, when g is modeled with a linear function, our test loses its validity, rejecting values of δ that are larger than the true bias. Such behavior is expected as the chosen function class $\mathcal{G}$ lacks the complexity necessary to capture the true bias model. Nevertheless, we observe that the *small* network with one hidden layer is already sufficient. Further, significantly increasing the complexity – the *large* network has approximately 45 times more parameters than the *small* one – still yields high power. Therefore, we recommend practitioners to be conservative in their choice of function class to ensure validity, even if it might come at the potential cost of some power and a more complex optimization problem. Moreover, although we cannot guarantee convergence to a global optimum, given the non-convexity of the problem for complex function classes, we show that the optimization procedure is stable and consistently reaches the same solution in Appendix C.3.

5 Real-world experiments

In this section, we provide a concrete application of the benchmarking framework using the Women's Health Initiative (WHI) study. In doing so, we show the strengths of our testing procedure and how tolerance and granularity are necessary for effective benchmarking.

5.1 The WHI controversy

The WHI was a collection of a randomized trial and an observational study that investigated the use of hormone therapy (HT) for preventing common sources of mortality among postmenopausal women, including cardiovascular disease, cancer, and fractures [2].

To HT, or not to HT The initial results of the WHI study in 2002 led to fear and confusion regarding the use of hormone therapy (HT) after menopause, resulting in a dramatic reduction in prescriptions for HT around the world. Although in 2002, it was stated that HT increases the risk of coronary heart disease (CHD) for all women, subsequent studies clearly showed that younger women close to menopause can benefit from HT. Indeed, for at least 2 decades before the WHI study, observational studies had suggested that HT reduces the risk of CHD [15, 16, 18, 48]. Further, subsequent randomized trials have continued demonstrating the benefits of HT when started early in young women close to menopause [21, 51]. Indeed, current epidemiological knowledge suggests that HT is expected to reduce the risk of CHD in women aged less than 60 years and within 10 years of menopause [32].

Limitations of the WHI randomized trial The main issue with the randomized trial from the WHI study is that younger women's cardiac events are relatively rare. Hence, the trial lacked enough events to reach statistical significance on these subgroups, and the average treatment effect suggested an increase in CHD risk because the majority of adverse events came from older women. Indeed, not only would it have been prohibitively expensive to conduct a randomized trial exclusively in younger women, but it would have also taken many years to accumulate enough events to reach statistical significance.

Benchmarking can help! It has been argued that in the 10 years since the WHI study, many women have been denied HT, significantly disadvantaging a generation of women [49]. The natural question is, thus, if, going back in time, benchmarking the observational study could have prevented such a turn of events. Indeed, this is the perfect setting to test our methodology, as we would like to ask the question:

Is the bias in the observational study enough to explain away
the benefits of HT in young women close to menopause?

In what follows, we show that answering such a question requires a statistical test that offers tolerance. Further, even though granularity is not required in this concrete example – false negatives are not possible since this particular subgroup is not heavily biased – we stress that it is equally important in practice. This is especially true w.r.t. to age and time since the start of menopause, as $\hat{\phi}^{\mathrm{ATE}}$ can fail to detect subgroup bias that cancels on average, as shown in our semi-synthetic experiments.

Table 1: The significance level is set at $\alpha = 0.05$. $\hat{\delta}_{\text{CT}}$ is the amount of bias that would explain away the positive effect of HT in young women close to menopause. $\hat{\delta}_{\text{LB}}$ is the smallest amount of bias for which the respective test starts accepting. $\hat{\phi}_0^{\text{ATE}}$ and $\hat{\phi}_0^{\text{CATE}}$ denote the respective tests when the tolerance function is set at $\delta = 0$.

Statistical tests	$\hat{\phi}^{\text{CATE}}$	$\hat{\phi}^{\text{ATE}}$	$\hat{\phi}_0^{\text{CATE}}$	$\hat{\phi}_0^{\text{ATE}}$
$\hat{\delta}_{\text{CT}}$	0.32	0.32	0.32	0.32
$\hat{\delta}_{\text{LB}}$	**0.25**	0.11	✗	✗
Reject the study	0	0	1	1

5.2 Experimental results

Linking back to our question of interest, we demonstrate how our method can provide a correct answer, i.e. one that aligns with the epidemiology literature. A natural way to do so is to first estimate from the available data the amount of bias that would explain away the treatment effect on the group of interest, defined as

$$\hat{\delta}_{\text{CT}} := \left| \mathbb{E}_{\mathbb{P}^{\text{os}}}[\tau^{\text{os}}(X) \mid X \in G] \right|.$$

In essence, the critical value quantifies the minimum strength of bias for which positive and negative values of treatment effect are reasonable, thereby invalidating the observational study results[3]. In our example, the group G is defined as young women (age ≤ 60) who are close to menopause (≤ 10 years).

Similarly to the semi-synthetic experiments, we instantiate the tolerance functions using constant upper and lower bounds, i.e. $\tau_{\pm}^{\text{os}}(X) = \tau^{\text{os}}(X) \pm \delta$ for some constant $\delta \in \mathbb{R}^+$. We compute the lower bound $\hat{\delta}_{\text{LB}}$ on the maximum amount of bias in the observational study, as defined in Equation (14). Note that this quantity can be computed only for tests that allow some tolerance. Then, our decision-making procedure will flag the observational study as biased if $\hat{\delta}_{\text{LB}} \geq \hat{\delta}_{\text{CT}}$.

Experimental details We consider a binary-valued outcome: the presence of coronary heart disease within the follow-up period. We choose as covariates X the basic adjustment variables used in many existing analyses, and we further limit patients to those who were not current users of HT at the time of enrolment, as the duration of HT use has been found to have a substantial impact on treatment effects [41, 52]. We refer to Appendix B.2 for complete experimental details.

We now present evidence that our procedure can yield the conclusions established in the epidemiological literature. In doing so, it avoids issuing false alarms when the bias is negligible (tolerance). Further, it detects a larger amount of bias, as it is more powerful than tests based on average treatment effect (granularity).

Results In Table 1, we show the result for all the statistical tests on the WHI study. First, we observe that both tests that allow for tolerance correctly do not flag the study, while $\hat{\phi}_0^{\text{CATE}}$ and $\hat{\phi}_0^{\text{ATE}}$ do. This difference shows the importance of tolerance for distinguishing between small and large amounts of bias. Second, we observe that the lower bound on the bias is larger for $\hat{\phi}^{\text{CATE}}$. Such behavior is expected and shows the importance of granularity to spot bias that would otherwise go unnoticed using $\hat{\phi}^{\text{ATE}}$.

[3]Note that other choices for the critical value are possible, and practitioners should determine the most appropriate one given the specific context.

6 Limitations and future work

Our approach shares limitations with other methods that rely on kernels for testing. Most notably, the curse of dimensionality can be a significant problem given the small sample size of randomized trials. In addition, the benchmarking strategy is optimistic; outside the common support of the two studies, the maximum bias could be arbitrarily high.

Our discussion suggests several important directions for future research. For example, our test could be adapted to the scenario where multiple observational datasets may be available but no randomized trials. Further, in settings where $n_{\text{rct}} \approx n_{\text{os}}$, or the tolerance functions $\tau_{\pm}^{\text{os}}$ are difficult to learn, Assumption (*iv*) in Theorem 3.1 may be unrealistic. One way to overcome this limitation is to construct a doubly robust test statistic that effectively combines multiple nuisance functions to relax the required assumptions on the approximation quality of the individual nuisance functions. Finally, applying and validating our method in more real-world scenarios presents an exciting avenue for future work.

Acknowledgements

PDB was supported by the Hasler Foundation grant number 21050. JA was supported by the ETH AI Center. KD was supported by the ETH AI Center and the ETH Foundations of Data Science.

References

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Garnet L Anderson, Joann Manson, Robert Wallace, Bernedine Lund, Dallas Hall, Scott Davis, Sally Shumaker, Ching-Yun Wang, Evan Stein, and Ross L Prentice. Implementation of the Women's Health Initiative study design. *Annals of Epidemiology*, 13(9):S5–S17, 2003.

[3] David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

[4] Yuwen Cheng, Lili Wu, and Shu Yang. Enhancing treatment effect estimation: A model robust approach integrating randomized experiments and external controls using the double penalty integration estimator. *Conference on Uncertainty in Artificial Intelligence*, 2023.

[5] Issa J Dahabreh, James M Robins, and Miguel A Hernán. Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology*, 31(5):614–619, 2020.

[6] Issa J Dahabreh, James M Robins, Sebastien JP Haneuse, Sarah E Robertson, Jon A Steingrimsson, and Miguel A Hernán. Global sensitivity analysis for studies extending inferences from a randomized trial to a target population. *arXiv preprint arXiv:2207.09982*, 2022.

[7] Issa J Dahabreh, James M Robins, Sebastien J-PA Haneuse, Iman Saeed, Sarah E Robertson, Elizabeth A Stuart, and Miguel A Hernán. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Statistics in Medicine*, 42(13):2029–2043, 2023.

[8] Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *International Conference on Artificial Intelligence and Statistics*, 2024.

[9] Miguel A Delgado. Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, 17(3):199–204, 1993.

[10] Ilker Demirel, Edward De Brouwer, Zeshan Hussain, Michael Oberst, Anthony Philippakis, and David Sontag. Benchmarking observational studies with experimental data under right-censoring. *International Conference on Artificial Intelligence and Statistics*, 2024.

[11] Narjust Duma, Jesus Vera Aguilera, Jonas Paludo, Candace L Haddox, Miguel Gonzalez Velez, Yucai Wang, Konstantinos Leventakos, Joleen M Hubbard, Aaron S Mansfield, Ronald S Go, et al. Representation of minorities and women in oncology clinical trials: review of the past 14 years. *Journal of Oncology Practice*, 14(1):e1–e10, 2018.

[12] Shaun P Forbes and Issa J Dahabreh. Benchmarking observational analyses against randomized trials: a review of studies assessing propensity score methods. *Journal of General Internal Medicine*, 35: 1396–1404, 2020.

[13] Jessica M Franklin, Robert J Glynn, David Martin, and Sebastian Schneeweiss. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clinical Pharmacology & Therapeutics*, 105(4):867–877, 2019.

[14] Chenyin Gao and Shu Yang. Pretest estimation in combining probability and non-probability samples. *Electronic Journal of Statistics*, 17(1):1492–1546, 2023.

[15] Deborah Grady, Susan M Rubin, Diana B Petitti, Cary S Fox, Dennis Black, Bruce Ettinger, Virginia L Ernster, and Steven R Cummings. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Annals of Internal Medicine*, 117(12):1016–1037, 1992.

[16] Francine Grodstein, JoAnn E Manson, Graham A Colditz, Walter C Willett, Frank E Speizer, and Meir J Stampfer. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Annals of Internal Medicine*, 133(12):933–941, 2000.

[17] Zhe He, Xiang Tang, Xi Yang, Yi Guo, Thomas J George, Neil Charness, Kelsa Bartley Quan Hem, William Hogan, and Jiang Bian. Clinical trial generalizability assessment in the big data era: a review. *Clinical and Translational Science*, 13(4):675–684, 2020.

[18] Brian E Henderson, Annlia Paganini-Hill, and Ronald K Ross. Decreased mortality in users of estrogen replacement therapy. *Archives of Internal Medicine*, 151(1):75–78, 1991.

[19] Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

[20] K Hillstrom. The MineThatData e-mail analytics and data mining challenge, 2008.

[21] Howard N Hodis, Wendy J Mack, Victor W Henderson, Donna Shoupe, Matthew J Budoff, Juliana Hwang-Levine, Yanjie Li, Mei Feng, Laurie Dustin, Naoko Kono, et al. Vascular effects of early versus late postmenopausal treatment with estradiol. *New England Journal of Medicine*, 374(13):1221–1231, 2016.

[22] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[23] Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate u-statistics. *The Annals of Statistics*, pages 1811–1823, 1993.

[24] Zeshan Hussain, Ming-Chieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. *International Conference on Artificial Intelligence and Statistics*, 2023.

[25] Zeshan M Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, 2022.

[26] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *International Conference on Machine Learning*, 2021.

[27] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 2018.

[28] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. *International Conference on Artificial Intelligence and Statistics*, 2019.

[29] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[30] Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross u-statistics. *Bernoulli*, 30 (1):683–711, 2024.

[31] David C Klonoff. The new FDA real-world evidence program to support development of drugs and biologics. *Journal of Diabetes Science and Technology*, 14(2):345–349, 2020.

[32] Sa Ra Lee, Moon Kyoung Cho, Yeon Jean Cho, Sungwook Chun, Seung-Hwa Hong, Kyu Ri Hwang, Gyun-Ho Jeon, Jong Kil Joo, Seul Ki Kim, Dong Ock Lee, et al. The 2020 menopausal hormone therapy guidelines. *Journal of Menopausal Medicine*, 26(2):69, 2020.

[33] Alex Luedtke, Marco Carone, and Mark J van der Laan. An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):75–99, 2019.

[34] Marco Morucci, Vittorio Orlandi, Harsh Parikh, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach to combining experimental and observational data. *arXiv preprint arXiv:2307.01449*, 2023.

[35] Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. *Conference on Uncertainty in Artificial Intelligence*, 2020.

[36] Natalie Neumeyer and Holger Dette. Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920, 2003.

[37] Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R Cole, and Elizabeth A Stuart. Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, pages 225–247, 2017.

[38] Trang Quynh Nguyen, Benjamin Ackerman, Ian Schmid, Stephen R Cole, and Elizabeth A Stuart. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PloS One*, 13(12):e0208795, 2018.

[39] Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. *International Conference on Machine Learning*, 2023.

[40] Richard Platt, Jeffrey S Brown, Melissa Robb, Mark McClellan, Robert Ball, Michael D Nguyen, and Rachel E Sherman. The FDA Sentinel Initiative—an evolving national resource. *New England Journal of Medicine*, 379(22):2091–2093, 2018.

[41] Ross L Prentice, Robert Langer, Marcia L Stefanick, Barbara V Howard, Mary Pettinger, Garnet Anderson, David Barad, J David Curb, Jane Kotchen, Lewis Kuller, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *American Journal of Epidemiology*, 162(5): 404–414, 2005.

[42] Jeffery S Racine, Jeffrey Hart, and Qi Li. Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, 25(4):523–544, 2006.

[43] Evan TR Rosenman, Guillaume Basse, Art B Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023.

[44] Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

[45] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

[46] Beth Schurman. The framework for FDA's real-world evidence program. *Applied Clinical Trials*, 28(4), 2019.

[47] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

[48] Meir J Stampfer and Graham A Colditz. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Preventive Medicine*, 20(1):47–63, 1991.

[49] DW Sturdee, A Pines, and International Menopause Society Writing Group. Updated ims recommendations on postmenopausal hormone therapy and preventive strategies for midlife health. *Climacteric*, 14(3):302–320, 2011.

[50] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[51] Hugh S Taylor, Aya Tal, Lubna Pal, Fangyong Li, Dennis M Black, Eliot A Brinton, Matthew J Budoff, Marcelle I Cedars, Wei Du, Howard N Hodis, et al. Effects of oral vs transdermal estrogen therapy on sexual function in early postmenopause: ancillary study of the kronos early estrogen prevention study (keeps). *JAMA Internal Medicine*, 177(10):1471–1479, 2017.

[52] Jan P Vandenbroucke. The HRT controversy: observational studies and RCTs fall in line. *The Lancet*, 373(9671):1233–1235, 2009.

[53] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014.

[54] Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. *Conference on Causal Learning and Reasoning*, 2022.

[55] Shu Yang, Donglin Zeng, and Xiaofei Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020.

[56] Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, 04 2023.

A Methodology

A.1 Power of the test

In this section, we discuss a simple result exemplifying the power of the test from Equation (10) in rejecting an alternative hypothesis. While Hussain et al. [24], Muandet et al. [35] show asymptotic normality for the kernel conditional moment test statistics $\mathbb{H}^2$ under the alternative hypothesis that $E\mathbb{H}^2 > 0$, this is no longer straightforwardly the case for the test statistics used in Equation (10) since it is a solution of an optimization problem. However, as we show in the following theorem, the statistic grows at a rate $\sqrt{n}$. For simplicity of the statement, we only prove the result for the oracle test statistics ψ and note that the generalization to approximate test statistics follows straightforwardly using the assumptions in Theorem 3.1 from the same argument as used in the proof of Theorem 3.1.

Theorem A.1. *Given the following two assumptions*

(i) the outcome variable Y and kernel k are uniformly bounded by constants

(ii) for every $\epsilon > 0$, the function class $\mathcal{G}$ has a finite ℓ_∞-norm covering number

Then, we can lower-bound the test statistics in probability as $n_{\text{rct}} \to \infty$ from Equation (10) by

$$\min_{g \in \mathcal{G}} \left| \frac{\sqrt{n_{\text{rct}}}\ \hat{\mathbb{H}}^2(\psi_g)}{\hat{\sigma}\left(\hat{\mathbb{H}}^2(\psi_g)\right)} \right| \gtrsim \sqrt{n_{\text{rct}}}\ \left(\inf_{g \in \mathcal{G}} \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_{\mathbb{P}^{\text{rct}}} \left[\psi_g(Z) f(X^{\mathcal{J}}) \right] \right)^2$$

where we use $\gtrsim$ to hide universal constants not depending on n_{rct}.

As we can see, under the alternative hypothesis, the RHS grows at a rate $\sqrt{n}$, where we hide in the constant dependencies on $\mathcal{G}$. Thus, our test statistics yield the same asymptotic power (up to constant factors) as standard kernel conditional moment test without tolerance, i.e. where $\tau_-^{\text{os}} = \tau_+^{\text{os}}$ (see [24, 35]).

Proof of Theorem A.1 The proof follows from applying a standard ϵ-net argument. Denote with $T := \inf_{g \in \mathcal{G}} \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_{\mathbb{P}^{\text{rct}}} [\psi_g(Z) f(V)]$ and note that if $T = 0$ the result follows trivially. Thus, we may assume that $T > 0$ is some constant independent of n (since we assume that ψ is fixed and does not change with n). First, note that by Assumption (i), we have that ψ is also uniformly bounded. Thus, it is straightforward to see that the variance term $\hat{\sigma}\left(\hat{\mathbb{H}}^2(\psi_g)\right)$ is also uniformly bounded. Hence, it suffices to show that $\hat{\mathbb{H}}^2(\psi_g) = \Omega_{\mathbb{P}}(1)$ is lower bounded in probability.

First note that for all $g \in \mathcal{G}$, we have that

$$\mathbb{E}\ \hat{\mathbb{H}}^2(\psi_g) = \mathbb{E}_{Z_1, Z_2} \psi_g(Z_1) k(X_1^{\mathcal{J}}, X_2^{\mathcal{J}}) \psi_g(Z_2) \geq \inf_{g \in \mathcal{G}} \mathbb{E}_{Z_1, Z_2} \psi_g(Z_1) k(X_1^{\mathcal{J}}, X_2^{\mathcal{J}}) = T_1^2 \tag{15}$$

where the last inequality follows from Equation (7). Thus, it suffices to show that the following inequality holds in probability

$$\sup_{g \in \mathcal{G}} |\hat{\mathbb{H}}^2(\psi_g) - \mathbb{E}\ \hat{\mathbb{H}}^2(\psi_g)| \leq \frac{T_1^2}{2}. \tag{16}$$

We use a simple, crude ϵ-net argument to show this result. Let $\mathcal{G}_\epsilon$ be the epsilon net in ℓ_∞ distance of balls of radii ϵ. That is $\forall g \in \mathcal{G},\ \exists \tilde{g} \in \mathcal{G}_\epsilon$ such that $\sup_{X^{\mathcal{J}} \in \text{supp}(\mathbb{P}^{\text{rct}}_{X^{\mathcal{J}}})} |g(X^{\mathcal{J}}) - \tilde{g}(X^{\mathcal{J}})| \leq \epsilon$. First, note that for

every fixed $\epsilon > 0$, by assumption $|\mathcal{G}_\epsilon| < \infty$ is finite. Thus, we can apply the law of large numbers to show that $\sup_{\tilde{g}\in\mathcal{G}_\epsilon} |\hat{\mathbb{H}}^2(\psi_{\tilde{g}}) - \mathbb{E}\,\hat{\mathbb{H}}^2(\psi_{\tilde{g}})| \to 0$. Hence, it suffices to show that we can choose a constant $\epsilon > 0$, such that the following inequality holds in probability,

$$\sup_{g\in\mathcal{G}} \inf_{\tilde{g}\in\mathcal{G}_\epsilon} |\hat{\mathbb{H}}^2(\psi_g) - \hat{\mathbb{H}}^2(\psi_{\tilde{g}})| \leq T^2/4. \tag{17}$$

Since ψ is uniformly bounded, we have that for all Z and $g \in \mathcal{G}$, $|\psi_g(Z) - \psi_{\tilde{g}}(Z)| \lesssim \epsilon$. Thus, we can use this fact in combination with the definition of $\hat{\mathbb{H}}^2(\psi_g)$ from Equation (8) to find $\epsilon > 0$ such that Equation (17) holds and thus conclude the proof.

A.2 Proof of Theorem 3.1

For the simplicity of notation, we write $n = n_{\text{rct}}$ throughout the proof. We begin the proof with the simple observation

$$\min_{g\in\mathcal{G}} \left| \frac{\sqrt{n}\hat{\mathbb{H}}^2(\hat{\psi}_g)}{\hat{\sigma}\left(\hat{\mathbb{H}}^2(\hat{\psi}_g)\right)} \right| \leq \left| \frac{\sqrt{n}\hat{\mathbb{H}}^2(\hat{\psi}_{g^*})}{\hat{\sigma}\left(\hat{\mathbb{H}}^2(\hat{\psi}_{g^*})\right)} \right| \tag{18}$$

which holds under the assumption that $g^* \in \mathcal{G}$. Thus, the statement in Theorem 3.1 follows when showing that the RHS converges in distribution to an absolute normal distribution. We prove this statement by showing that

$$n\hat{\mathbb{H}}^2(\hat{\psi}_{g^\star}) \to n\hat{\mathbb{H}}^2(\psi_{g^\star}) \quad \text{and} \quad n\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\hat{\psi}_{g^\star})\right) \to n\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\psi_{g^\star})\right). \tag{19}$$

Using Assumptions (i),(ii), and (iii) in the theorem statement, we can then apply [30] Theorem 4.2 to show that

$$\frac{\sqrt{n}\hat{\mathbb{H}}^2(\psi_{g^*})}{\hat{\sigma}\left(\hat{\mathbb{H}}^2(\psi_{g^*})\right)} \to \mathcal{N}(0,1). \tag{20}$$

Moreover, as a consequence of Equation (12) and (57) in the Proof of Theorem 4.2 in [30], we further have that

$$\frac{1}{n\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\psi_{g^\star})\right)} = O_{\mathbb{P}}(1). \tag{21}$$

Thus, when applying Slutsky's Theorem we conclude that

$$\frac{\sqrt{n}\hat{\mathbb{H}}^2(\psi_{g^*})}{\hat{\sigma}\left(\hat{\mathbb{H}}^2(\psi_{g^*})\right)} \to N(0,1), \tag{22}$$

and thus also the statement in Theorem 3.1.

A.3 Proof of statement in Equation 19

For simplicity of notation, throughout the rest of the proof, we generally write $\hat{\psi}_i := \hat{\psi}_{g^\star}(Z_i)$ and $\psi_i := \psi_{g^\star}(Z_i)$. We define

$$\triangle := \hat{\psi} - \psi = g(X)\left(\hat{\tau^{\text{os}}}_+ - \tau^{\text{os}}_+\right) + (1 - g(X))\left(\hat{\tau^{\text{os}}}_- - \tau^{\text{os}}_-\right). \tag{23}$$

and we write $\triangle_i := \triangle(Z_i)$. Further, we use the convention that $I \in \mathcal{I}_1$ and $J \in \mathcal{I}_2$ are two arbitrarily chosen indices from the corresponding index stets. Moreover, let us restate the definition of the mean and variance terms, as defined in Section 3.1 and Section 3.2

$$n\hat{\mathbb{H}}^2(\hat{\psi}) = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_1} \hat{\psi}_i \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{I}_2} k(X_i^{\mathcal{J}}, X_j^{\mathcal{J}}) \hat{\psi}_j \tag{24}$$

$$n\hat{\sigma}^2 \left(\hat{\mathbb{H}}^2(\hat{\psi}) \right) = \frac{1}{n} \sum_{i \in \mathcal{I}_1} \hat{\psi}_i^2 \left(\frac{1}{\sqrt{n}} \sum_{j \in \mathcal{I}_2} k(X_i^{\mathcal{J}}, X_j^{\mathcal{J}}) \hat{\psi}_j \right)^2 - (\sqrt{n}\hat{\mathbb{H}}^2(\hat{\psi}))^2. \tag{25}$$

Preliminary step: bounds for the error term $\triangle_i$ By Assumption (iv) in Theorem 3.1, we have that

$$\mathbb{E}\triangle_I^2 =: \|\triangle\|_{L_2(\mathbb{P})}^2 \leq 2\|\tau^{\hat{\text{os}}}{}_+ - \tau_+^{\text{os}}\|_{L_2(\mathbb{P})}^2 + 2\|\tau^{\hat{\text{os}}}{}_- - \tau_-^{\text{os}}\|_{L_2(\mathbb{P})}^2 = o_{\mathbb{P}}\left(\frac{1}{n}\right), \tag{26}$$

where the probability is over the dataset $\mathcal{D}^{\text{os}}$ used to train $\hat{\tau}_{\pm}^{\text{os}}$. We further define

$$\tau_2(X_i^{\mathcal{J}}) := \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{I}_2} k(X_i^{\mathcal{J}}, X_j^{\mathcal{J}}) \triangle_j \quad \text{and} \quad \tau_1(X_j^{\mathcal{J}}) := \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_1} k(X_i^{\mathcal{J}}, X_j^{\mathcal{J}}) \triangle_i. \tag{27}$$

We will repetitively make use of the following two bounds, which hold analogously for τ_1:

$$\sup_{X^{\mathcal{J}}} |\mathbb{E}\left[\tau_2(X^{\mathcal{J}})|X^{\mathcal{J}}\right]| = \sup_{X^{\mathcal{J}}} |\sqrt{n}\mathbb{E}\left[k(X^{\mathcal{J}}, X_J^{\mathcal{J}})\triangle_J | X^{\mathcal{J}}\right]| \lesssim \sqrt{n}\sqrt{\mathbb{E}\triangle_J^2}, \tag{28}$$

where we used Cauchy-Schwartz in the last inequality and the fact that the kernel is uniformly bounded. Further, we use that

$$\sup_{X^{\mathcal{J}}} \left[\mathbb{E}\left[\left(\tau_2(X^{\mathcal{J}})\right)^2 | X^{\mathcal{J}}\right]\right] = \sup_{X^{\mathcal{J}}} \left[\mathbb{E}\, \frac{1}{n}\sum_j k(X^{\mathcal{J}}, X_j^{\mathcal{J}})^2 \triangle_j^2 + \frac{1}{n}\sum_{j \neq j'} k(X^{\mathcal{J}}, X_j^{\mathcal{J}}) k(X^{\mathcal{J}}, X_{j'}^{\mathcal{J}}) \triangle_j \triangle_{j'} | X^{\mathcal{J}}\right] \tag{29}$$

$$\lesssim \mathbb{E}\triangle_J^2 + \sup_{X^{\mathcal{J}}} \left[\frac{n(n-1)}{n} \left(\mathbb{E}[k(X^{\mathcal{J}}, X_J^{\mathcal{J}})\triangle_J | X^{\mathcal{J}}]\right)^2\right] \tag{30}$$

$$\leq \mathbb{E}\triangle_J^2 + (n-1) \sup_{X^{\mathcal{J}}} \left[\left(\mathbb{E}(k(X^{\mathcal{J}}, X_J^{\mathcal{J}}))^2 | X^{\mathcal{J}}\right) \left(\mathbb{E}\triangle_J^2\right)\right] = o_{\mathbb{P}}(1), \tag{31}$$

where we used Cauchy-Schwartz again. We can now bound both terms in Equation (19).

Term 1: controlling $n\hat{\mathbb{H}}^2(\hat{\psi})$ We first control the mean term $n\hat{\mathbb{H}}^2(\hat{\psi})$. Our goal is to show that

$$\left| n\hat{\mathbb{H}}^2(\hat{\psi}) - n\hat{\mathbb{H}}^2(\psi) \right| = o_{\mathbb{P}}(1). \tag{32}$$

We decompose the difference into the following three terms:

$$n\hat{\mathbb{H}}^2(\hat{\psi}) - n\hat{\mathbb{H}}^2(\psi) = \underbrace{\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_1} \psi_i \tau_2(X_i^{\mathcal{J}})}_{=:T_1} + \underbrace{\frac{1}{\sqrt{n}} \sum_{j \in \mathcal{I}_2} \psi_i \tau_1(X_j^{\mathcal{J}})}_{=:T_2} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_1} \triangle_i \tau_2(X_i^{\mathcal{J}})}_{=:T3} \tag{33}$$

To control the first two terms, we note that under the null hypothesis in Equation (5), we have that for all $X_i^{\mathcal{J}}$, $\mathbb{E}\left[\psi_i | X_i^{\mathcal{J}}\right] = 0$. Thus, it suffices to show that the variance goes to zero:

$$\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_1}\psi_i\tau_2(X_i^{\mathcal{J}})\right) = \mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_1}\psi_i\tau_2(X_i^{\mathcal{J}})\right)^2 \tag{34}$$

$$= \mathbb{E}\quad \mathbb{E}[\psi_I^2|X_I^{\mathcal{J}}]\quad \left(\tau_2(X_I^{\mathcal{J}})\right)^2 \lesssim \mathbb{E}\ \left(\tau_2(X_I^{\mathcal{J}})\right)^2 = o_{\mathbb{P}}(1), \tag{35}$$

where we used that as a consequence of Assumption (i) in Theorem 3.1 the conditional fourth moment of ψ is uniformly bounded, and thus also the conditional second moment $\mathbb{E}_Z[\psi^2(Z)|X^{\mathcal{J}}]$ is uniformly bounded. Since the same argument also applies when swapping $\mathcal{I}_1$ with $\mathcal{I}_2$, we can conclude from Chebyshev's inequality that

$$|T_1| = o_{\mathbb{P}}(1) \quad \text{and} \quad |T_2| = o_{\mathbb{P}}(1). \tag{36}$$

Next, we bound the last term T_3. We separately consider the mean and variance of T_3:

$$\mathbb{E}\ T_3 = \sqrt{n}\ \mathbb{E}\ \triangle_I \tau_2(X_I^{\mathcal{J}}) = \sqrt{n}\ \mathbb{E}\ \triangle_I\ \mathbb{E}[\tau_2(X_I^{\mathcal{J}})|X_I^{\mathcal{J}}] \tag{37}$$

$$\leq \sup_{X^{\mathcal{J}}}\left[|\mathbb{E}\left[\tau_2(X^{\mathcal{J}})|X^{\mathcal{J}}\right]|\right]\mathbb{E}\left[\sqrt{n}|\triangle_I|\right] \leq \sup_{X^{\mathcal{J}}}\left[|\mathbb{E}\left[\tau_2(X^{\mathcal{J}})|X^{\mathcal{J}}\right]|\right]\sqrt{n}\left(\mathbb{E}\triangle_I^2\right)^{1/2} = o_{\mathbb{P}}(1) \tag{38}$$

and

$$\mathbb{E}\ T_3^2 = \mathbb{E}\ \triangle_I^2\left(\tau_2(X_I^{\mathcal{J}})\right)^2 = \mathbb{E}\ \triangle_I^2\ \mathbb{E}\left[\left(\tau_2(X_I^{\mathcal{J}})\right)^2|X_I^{\mathcal{J}}\right] \tag{39}$$

$$\leq \sup_{X^{\mathcal{J}}}\left[\mathbb{E}\left[\left(\tau_2(X^{\mathcal{J}})\right)^2|X^{\mathcal{J}}\right]\right]\mathbb{E}\ \triangle_I^2 = o_{\mathbb{P}}(1) \tag{40}$$

and thus we can further conclude that $|T_3| = o_{\mathbb{P}}(1)$.

Term 2: controlling $\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\hat{\psi})\right)$ As a second step, we control the variance term $\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\hat{\psi})\right)$. Our goal is again to show that

$$\left|n\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\hat{\psi})\right) - n\hat{\sigma}^2\left(\hat{\mathbb{H}}^2(\psi)\right)\right| = o_{\mathbb{P}}(1). \tag{41}$$

Given the results from the previous paragraph in Equation (32), we note that it suffices to show that

$$\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\hat{\psi}_i^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\hat{\psi}_j\right)^2 - \frac{1}{n}\sum_{i\in\mathcal{I}_1}\psi_i^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2\right| = o_{\mathbb{P}}(1). \tag{42}$$

We begin again by decomposing the difference of the two terms on the LHS into the following six terms:

$$= \underbrace{\frac{1}{n}\sum_{i\in\mathcal{I}_1}\triangle_i^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})(\psi_j+\triangle_j)\right)^2}_{=:T_1} + \underbrace{\frac{1}{n}\sum_{i\in\mathcal{I}_1}(\psi_i+\triangle_i)^2\left(\tau_2(X_i^{\mathcal{J}})\right)^2}_{=:T_2} - \underbrace{\frac{1}{n}\sum_{i\in\mathcal{I}_1}\triangle_i^2\left(\tau_2(X_i^{\mathcal{J}})\right)^2}_{=:T_3}$$

$$+ \underbrace{\frac{2}{n}\sum_{i\in\mathcal{I}_1}\psi_i^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)\tau_2(X_i^{\mathcal{J}})}_{=:T_4} + \underbrace{\frac{4}{n}\sum_{i\in\mathcal{I}_1}\psi_i\triangle_j\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)\tau_2(X_i^{\mathcal{J}})}_{=:T_5}$$

$$+ \underbrace{\frac{2}{n}\sum_{i\in\mathcal{I}_1}\psi_i\triangle_j\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2}_{=:T_6}$$

We now show for each term individually $\forall i \in \{1,\cdots,6\}, |T_i| = o_{\mathbb{P}}(1)$.

Controlling T_1: Since the term is non-negative, it suffices to show that the expectation $\mathbb{E}T_1 \to 0$ and apply Markov's inequality. We have

$$\begin{aligned}\mathbb{E}\ T_1 &= \mathbb{E}\ \triangle_I^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})(\psi_j+\triangle_j)\right)^2 = \mathbb{E}\ \triangle_I^2\Big[\frac{1}{n}\sum_{j\in\mathcal{I}_2}k^2(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j^2 + \left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\triangle_j\right)^2 \\ &+ \frac{1}{n}\sum_{j\in\mathcal{I}_2}k^2(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\triangle_j\Big] = \mathbb{E}\ \triangle_I^2 k(X_I^{\mathcal{J}},X_J^{\mathcal{J}})^2[\psi_J^2+\psi_J\triangle_J] + \mathbb{E}\ \triangle_I^2\left(\tau_2(X_I^{\mathcal{J}})\right)^2 \qquad (43)\\ &\lesssim \left(\mathbb{E}\triangle_I^2\right)\left[\mathbb{E}\ [\psi_J^2] + \left(\mathbb{E}[\psi_J^2]\mathbb{E}[\triangle_J^2]\right)^{1/2} + \sup_{X^{\mathcal{J}}}\mathbb{E}\left[\left(\tau_2(X^{\mathcal{J}})\right)^2|X^{\mathcal{J}}\right]\right] = o_{\mathbb{P}}(1)\end{aligned}$$

where we used in the second equality that $\mathbb{E}\ [\psi_j|X_j^{\mathcal{J}}] = 0$.

Controlling T_2 and T_3: We can again bound the expectation and apply Markov's inequality. We have

$$\mathbb{E}\ T_2 \le \mathbb{E}\left[\left(2\psi_i^2+2\triangle_i^2\right)\mathbb{E}\left[\ \left(\tau_2(X_i^{\mathcal{J}})\right)^2|X_i^{\mathcal{J}}\right]\right] = \sup_{X^{\mathcal{J}}}\left(\mathbb{E}\left[\left(\tau_2(X^{\mathcal{J}})\right)^2|X^{\mathcal{J}}\right]\right)\left(2\mathbb{E}\ \psi_I^2+2\mathbb{E}\triangle_I^2\right) = o_{\mathbb{P}}(1) \quad (44)$$

and thus also $\mathbb{E}\ T_3 = o(1)$.

Controlling T_4, T_5 and T_6: We note that the expectations $\mathbb{E}T_4 = 0$, $\mathbb{E}T_5 = 0$ and $\mathbb{E}T_6 = 0$ are all zero. Thus, we can bound the terms in probability by showing that the Variance converges to zero and applying Chebyshev's inequality. We make use of the fact that the fourth conditional moment of ψ given $X^{\mathcal{J}}$ is almost surely upper bounded by a constant, which follows from the fact that the conditional moment of Y given $X^{\mathcal{J}}$ is uniformly upper bounded (Assumption (i) in Theorem 3.1). We can, therefore, bound the variance of T_4 by

$$\mathrm{Var}\left(\frac{2}{n}\sum_{i\in\mathcal{I}_1}\psi_i^2\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)\tau_2(X_i^{\mathcal{J}})\right) = \frac{4}{n}\mathbb{E}\ \psi_I^4\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2\left(\tau_2(X_I^{\mathcal{J}})\right)^2 \quad (45)$$

$$=4\mathbb{E}\ \psi_I^4\left(\frac{1}{n}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2\left(\tau_2(X_I^{\mathcal{J}})\right)^2 \lesssim \sup_{X^{\mathcal{J}}}\mathbb{E}[\left(\tau_2(X^{\mathcal{J}})\right)^2|X^{\mathcal{J}}] = o_{\mathbb{P}}(1), \quad (46)$$

where J is again any arbitrary index in $\mathcal{I}_2$ and we used that the fourth conditional moments of $\mathbb{E}\psi^4|X$ is uniformly upper bounded, as well as the kernel k. Next, we bound the variance of T_5:

$$\text{Var}\left(\frac{2}{n}\sum_{i\in\mathcal{I}_1}\psi_i\triangle_j\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)\tau_2(X_i^{\mathcal{J}})\right) \tag{47}$$

$$=4\mathbb{E}\ \psi_I^2\triangle_I^2\left(\frac{1}{n}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2\left(\tau_2(X_I^{\mathcal{J}})\right)^2=o_{\mathbb{P}}(1) \tag{48}$$

and finally, the variance of the term T_6:

$$\text{Var}\left(\frac{2}{n}\sum_{i\in\mathcal{I}_1}\psi_i\triangle_i\left(\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{I}_2}k(X_i^{\mathcal{J}},X_j^{\mathcal{J}})\psi_j\right)^2\right) \tag{49}$$

$$=\ \frac{4}{n}\mathbb{E}\ \psi_I^2\triangle_I^2\left(\frac{1}{n^2}\sum_{j\in\mathcal{I}_2}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})^4\psi_j^4+\frac{6}{n^2}\sum_{j,j'\in\mathcal{I}_2;j\neq j'}k(X_I^{\mathcal{J}},X_j^{\mathcal{J}})^2k(X_I^{\mathcal{J}},X_{j'}^{\mathcal{J}})^2\psi_j^2\psi_{j'}^2\right)=o_{\mathbb{P}}(1) \tag{50}$$

As a result, we conclude that $|T_4|=o_{\mathbb{P}}(1)$ and $|T_5|=o_{\mathbb{P}}(1)$ and $|T_6|=o_{\mathbb{P}}(1)$ and thus conclude the second step and therefore the proof.

B Experimental details

B.1 Semi-synthetic experiments

Hillstrom's MineThatData Email dataset [20] is a large-scale, real-world randomized trial that contains records of 64,000 customers who made purchases online within the last twelve months. They were part of an email campaign designed to assess the effectiveness of different campaign strategies. Two treatment groups, "Men's" and "Women's" email campaigns, and a control group were established, with treatments randomly assigned. Our analysis primarily focuses on a combined treatment group, which constitutes approximately 66% of the dataset. Although the original dataset presents various outcomes, including binary indicators of customers visiting or purchasing in the days after the campaign, we focus on the dollars spent in the two weeks post-campaign. The dataset provides data on annual spending (History), merchandise type (Mens or Womens), geographical location (Zip Code), newcomer status (Newbie), and purchasing avenues (Channel). We, therefore, discard features describing the history segment (History segment) and recency of the last purchase (Recency). Since the average treatment effect is close to zero, we add a constant shift of 30 to all treated individuals, allowing us more flexibility to introduce bias. We normalize continuous features and one-hot-encode categorical features, resulting in a 13-dimensional dataset. By default, we use 80% of the full dataset as the observational study (os), and the remaining 20% as the randomized controlled trial (rct).

We fit the propensity score using logistic regression with default hyperparameters from `scikit-learn`. We train a `Random Forest Classifier` for the selection score (rct or os), also with default hyperparameters from `scikit-learn`. Finally, we estimate the CATE functions using the doubly-robust learner from Kennedy [29], instantiating `Random Forest Regressors` for the potential outcome functions and the pseudo-outcome regression, fixing hyperparameters to 300 `tree estimators` with a `maximum depth` of 6.

We sample the coefficient for the polynomial bias model in Figure 1b from a normal distribution $\mathcal{N}(0, 0.01^2)$.

B.2 Women's Health Initiative

The Women's Health Initiative (WHI) is a long-term national health study that has focused on strategies for preventing the major causes of death, disability, and frailty in older women, specifically heart disease, cancer, and osteoporotic fractures. This multi-million dollar, 20+ year project, sponsored by the National Institutes of Health (NIH), the National Heart, Lung, and Blood Institute (NHLBI), initially enrolled 161,808 women aged 50-79 between 1993 and 1998. The WHI was one of the most definitive, far-reaching clinical trials of post-menopausal women's health ever undertaken in the US.

The WHI had two major parts: a Clinical Trial and an Observational Study. The randomized controlled Clinical Trial (CT) enrolled 68,132 women in trials testing three prevention strategies. Eligible women could choose to enroll in one, two, or three of the trial components.

- Hormone Therapy Trials (HT): This component examined the effects of combined hormones or estrogen alone on the prevention of heart disease and osteoporotic fractures and associated risk for breast cancer. Women participating in this component took hormone pills or a placebo (inactive pill) until the Estrogen plus Progestin and Estrogen Alone trials were stopped early in July 2002 and March 2004, respectively. All HT participants continued to be followed without intervention until close-out.
- Dietary Modification Trial (DM): The Dietary Modification component evaluated the effect of a low-fat and high-fruit, vegetable, and grain diet on preventing breast and colorectal cancers and heart disease. Study participants followed either their usual eating pattern or a low-fat dietary pattern.

- Calcium/Vitamin D Trial (CaD): This component evaluated the effect of calcium and vitamin D supplementation on preventing osteoporotic fractures and colorectal cancer. Women in this component took calcium and vitamin D pills or placebos.

The Observational Study (OS) examines the relationship between lifestyle, health and risk factors and disease outcomes. This component involves tracking the medical events and health habits of 93,676 women. Recruitment for the observational study was completed in 1998, and participants have been followed since.

We use observational study and randomized trial data from the Women's Health Initiative (WHI) to assess our method in a real-world scenario. We use the Postmenopausal Hormone Therapy (PHT) trial as the RCT in our analysis ($n_{\text{rct}} = 16,608$), 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 outcomes were tracked similar to those of the RCT.

Data preprocessing We binarize a composite outcome, where $Y = 1$ if coronary heart disease was observed in the first seven years of follow-up, and $Y = 0$ otherwise. 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. Using this procedure, we end up with a total of $n_{\text{os}} = 33,511$ patients. Finally, we restrict the set of covariates used to those that are measured in both the RCT and the observational study. In particular, we use as covariates only those measured in both the RCT and observational study, and we further restrict them to those identified as significant in epidemiological literature, such as in [41]. Specifically, the covariates in our analysis are: `AGE`, `ETHNIC_White`, `BMI`, `SMOKING_Past_Smoker`, `SMOKING_Current_Smoker`, `EDUC_x_College_graduate_or_Baccalaureate Degree`, `EDUC_x_Some_post-graduate_or_professional`, `MENO`, `PHYSFUN`. The data used is available on BIOLINCC.

Experimental details We use a gaussian kernel with `bandiwidth` = 1.0. We use a logistic regression model for both the outcome model and propensity score, default hyperparameters in `scikit-learn` were used. We train a neural (1 hidden layer and 10 neurons) network with Adam, with a learning rate of 0.01 for 500 epochs, we repeat the optimization for 10 seeds with different initialization to sanity check that we converge.

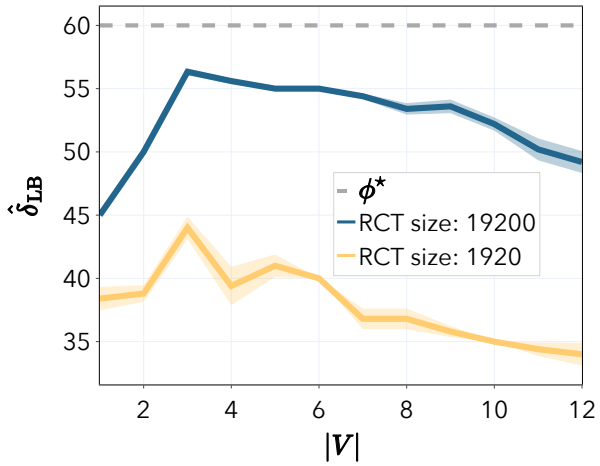


Figure 3: Effect of varying the feature set $X^{\mathcal{J}}$ on the average lower bound $\hat{\delta}_{\text{LB}}$ in scenario 1, illustrating the trade-off between feature set size and the power of the test. The highest power is achieved when the feature set size $|X^{\mathcal{J}}| = 3$, encompassing only the relevant features. The significance level is set at $\alpha = 0.05$, and $\phi^\star$ represents the oracle test, which rejects for $\delta < \delta^*$. We average runs over 5 seeds and report the standard error.

C Additional experiments

C.1 Ablations for the selected subset of features

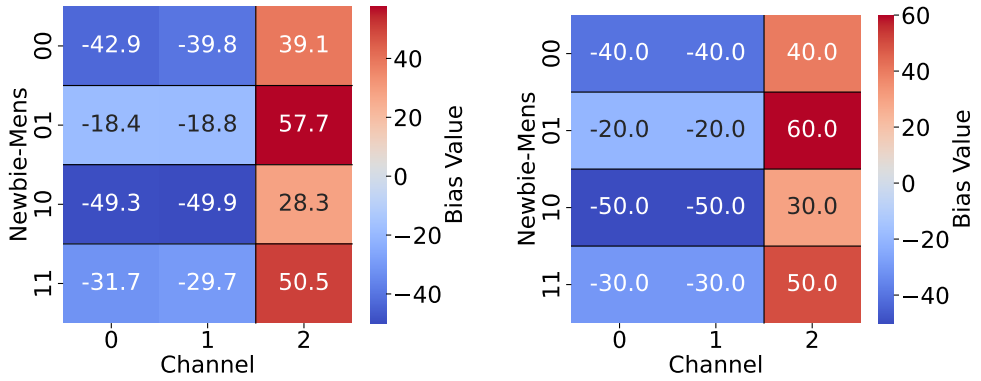
Figure 3 shows the effect of the selected feature set $X^{\mathcal{J}}$ on the average lower bound $\hat{\delta}_{\text{LB}}$ found by our testing procedure for the bias model from scenario 1 (see Figure 1a). In scenario 1, we introduced constant bias in the subgroups resulting from different combinations of the features "newbie", "mens" and "channel", with a maximum true bias $\delta^* = 60$. When $|X^{\mathcal{J}}| = 3$, we precisely select all features responsible for heterogeneity between rct and os datasets, achieving the highest power. Intuitively, the bias function cannot be accurately learned with smaller feature sets, and the test loses power. On the other hand, when increasing the feature set, the test also progressively loses power due to the curse of dimensionality, being particularly severe with smaller sample sizes. We note that we do not consider one-hot-encoded features in the visualization, i.e. when we choose all features $|X^{\mathcal{J}}| = 6$, we represent each data point with a 13-dimensional vector. After the sixth feature, we simply add redundant features sampled from a normal $\mathcal{N}(0, 1)$ distribution.

C.2 Interpretability of our testing procedure

Similar to the test proposed by Hussain et al. [24], our testing procedure also outputs a "witness function" that practitioners can leverage to identify the most biased subgroups in the observational dataset. In our case, it also hints at each subgroup's bias strength and direction. This is achieved by minimizing the objective in Equation (9), where we learn the function $g \in \mathcal{G}$. If the function class $\mathcal{G}$ is sufficiently rich to model the bias structure accurately, it effectively learns the discrepancy for each point in the rct necessary to fail to reject the null hypothesis of equality of CATEs between rct and os. Specifically, $g \in [0, 1]$ linearly interpolates between the tolerance bounds $\tau_-^{\text{os}}(X)$ and $\tau_+^{\text{os}}(X)$, so values close to zero suggest a negative bias of magnitude close to the estimated lower bound $\hat{\delta}_{\text{LB}}$, while values close to one indicate the same for positive bias.

This method allows a practitioner to estimate the subgroup bias as $\hat{bias}_G = \hat{\delta}\left(\frac{2}{n_G}\sum_{X_i \in G} g\left(X_i\right) - 1\right)$, where G represents the subgroup of interest and $\hat{\delta}$ the user shift corresponding to the learned g for which we fail to reject the null hypothesis, e.g. $\hat{\delta}_{\text{LB}}$. By examining specific subgroups, practitioners can "quantify" their bias, e.g. due to unobserved confounding or transportability violations, thus facilitating a better understanding of the bias structure. Figure 4 illustrates how practitioners could use the witness function for scenario 1 (Figure 1a), where the categorical nature of the features determines subgroups. We compare the estimated bias with the ground truth and observe that our estimation accurately reflects the true bias model. In scenarios where subgroups are not predefined, a practitioner could define them based on domain knowledge or select, for example, the bottom or top 10% of witness function values, as suggested by Hussain et al. [24].

However, it is crucial to note that, unlike the approach by Hussain et al. [24], we do not have guarantees for the correctness of the witness function since we rely on an optimization process. Therefore, any claims based on it should be approached cautiously and contrasted with domain expertise.



(a) Estimated bias model

(b) True bias model

Figure 4: Comparison between the estimated and true bias models for scenario 1, demonstrating the capability of the witness function to reflect the bias structure within subgroups accurately. We run the test with a random seed, using the same hyperparameters as in our experimental evaluation and choosing $\hat{\delta} = 57$.

C.3 Optimization convergence

We assess the reliability of our testing procedure by examining the behavior of the optimization process for Scenario 1, specifically during the training of the small neural network. Recall that, given the non-convex nature of the optimization problem, we cannot guarantee convergence to the true global minimum g^*. We plot in Figure 5 the test statistic with respect to the training epochs under different random network initializations, setting $\hat{\delta} = 58$. We observe that the test statistic consistently reaches the same local minimum and that the optimization process stabilizes after 10,000 epochs.

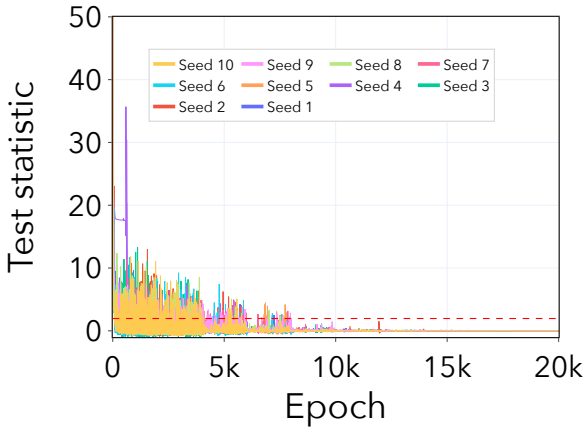


Figure 5: Evolution of the test statistic with respect to the training epochs for Scenario 1, during the training of the small neural network. We set $\hat{\delta} = 58$ ($\delta^* = 60$) and the significance level $\alpha = 0.05$. The test statistic corresponding to the fixed significance level α is also plotted. The rest of the hyperparameters remain the same as those described in the main paper.