Hidden yet quantifiable: A lower bound for confounding strength using randomized trials

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

Department of Computer Science, ETH Zürich

Abstract

In the era of fast-paced precision medicine, observational studies play a major role in properly evaluating new treatments in clinical practice. Yet, unobserved confounding can significantly compromise causal conclusions drawn from non-randomized data. We propose a novel strategy that leverages randomized trials to quantify unobserved confounding. First, we design a statistical test to detect unobserved confounding above a certain strength. Then, we use the test to estimate an asymptotically valid lower bound on the unobserved confounding strength. We evaluate the power and validity of our statistical test on several synthetic and semi-synthetic datasets. Further, we show how our lower bound can correctly identify the absence and presence of unobserved confounding in a real-world example.[1]

1 Introduction

Monitoring the performance of a newly approved treatment is crucial, a process commonly referred to as *post-marketing surveillance* [62]. Nowadays, the U.S. Food and Drug Administration promotes the integration of observational data in this process to address the shortcomings of randomized evidence [39, 48]. This strategy is essential for validating personalized treatments, like immunotherapy for certain types of cancer, where randomized evidence is scarce, and treatment costs are substantial [23, 25].

Yet, unobserved confounding can significantly compromise causal conclusions drawn from observational data. To tackle this issue, sensitivity analysis has been the prevalent paradigm since its conception by Cornfield et al. [10]. This field studies how a specific strength of unobserved confounding affects causal conclusions and introduces the concept of a *critical value* [33, 59], i.e. the minimum strength unobserved confounders would need to have to explain away the estimated treatment effect. However, critical values are solely based on observational data and can differ substantially from the *true confounding strength*. As a result, epidemiologists often rely on heuristic judgments to decide whether an observational study is flawed.

Estimating the true confounding strength is infeasible without further assumptions. Yet, once a treatment gains approval, we may have access to a randomized trial that allows for more effective strategies to address unobserved confounding. A recent line of works proposes to combine the estimators from randomized and observational data, e.g. see Brantner et al. [3], Colnet et al. [7] for a survey. However, these methods crucially rely on some prior knowledge of the confounding bias structure, that is not always available in practice.

*These authors contributed equally.

[1]See our GitHub repository for the source code: `https://github.com/jaabmar/confounder-lower-bound`.

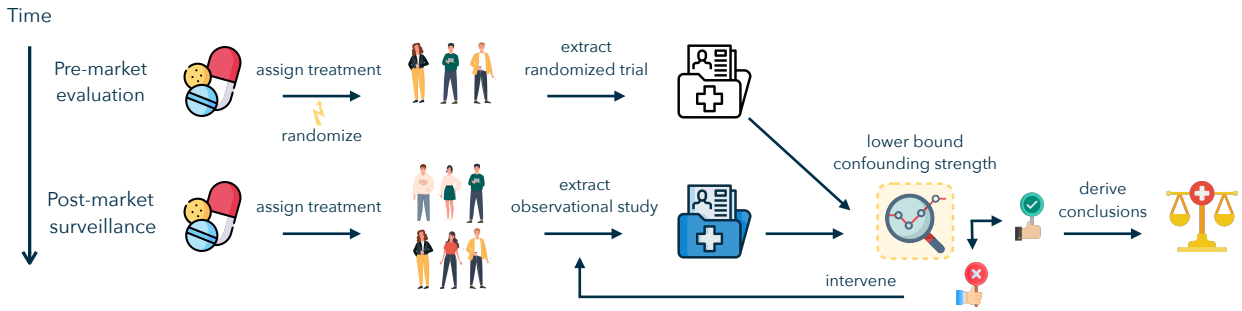


Figure 1: An illustrative example of the drug regulatory process: our lower bound allows taking proactive measures to address the unobserved confounding problem.

We propose an alternative strategy to leverage randomized trials, that is, to test and quantify the true confounding strength. In particular, if strong confounding is detected, epidemiologists can take proactive measures to correct it. Most directly, they can identify and incorporate relevant covariates into the study design if they were initially overlooked [19]. On the other hand, if small confounding is detected, epidemiologists can continue their analysis (see Figure 1 for an illustration of the pipeline). More concretely, our contributions are as follows.

- In Section 3, we introduce the first statistical test to detect unobserved confounding above a certain strength. Further, we show how the test can be used to estimate an asymptotically valid lower bound on the true confounding strength.
- In Section 4, we evaluate the finite-sample validity and power of our test on several synthetic and semi-synthetic datasets.
- In Section 5, we showcase through a real-world example how our approach leads to conclusions that align with established epidemiological knowledge.

1.1 Related work

Our approach is closely related to a line of work that proposes statistical tests for the *presence* of unobserved confounding. In particular, several works leverage randomized trials to detect unobserved confounding. These tests check for significant differences between average treatment effect estimates obtained from randomized and observational data [30, 42, 61, 65]. More sophisticated approaches also test for differences in conditional average treatment effect estimates [31] and account for right-censored outcomes [15].

Similarly, other works have designed statistical tests using instrumental variables and negative control outcomes instead of randomized trials [12, 16, 40, 53]. Additionally, multiple observational studies can be leveraged to test conditional independences and detect unobserved confounding [36].

In contrast to our test, these works have a significant limitation: they cannot quantify the true confounding strength. Even in infinite samples, they reject observational studies with negligible confounding. In real-world settings, where some degree of confounding will likely be present, testing for the absence of unobserved confounding can be too restrictive.

Finally, another line of works proposes calibrating the value of confounding strength using only observational data [29, 60]. However, the true confounding strength can be arbitrarily different from the calibrated strength, and there is no theoretical result for how these two quantities are related, even with infinite samples.

2 Setting and notation

We have access to data from a randomized trial (rct) and an observational study (os), which come from an underlying distribution $\mathbb{P}^{\diamond}_{\text{full}}$ over $(X, U, Y(0), Y(1), Y, T)$, for $\diamond \in \{\text{rct}, \text{os}\}$. Here, $(X, U) \in \mathbb{R}^d \times \mathbb{R}^k$ is a vector of confounders, $(Y(0), Y(1))$ are real-valued bounded potential outcomes, $Y \in \mathbb{R}$ is the observed outcome, and $T \in \{0, 1\}$ is a binary treatment indicator. However, the confounder U and the potential outcomes are never observed, that is, we can only sample from the distributions $\mathbb{P}^{\text{rct}} := \mathcal{M}(\mathbb{P}^{\text{rct}}_{\text{full}})$ and $\mathbb{P}^{\text{os}} := \mathcal{M}(\mathbb{P}^{\text{os}}_{\text{full}})$, where $\mathcal{M}(\mathbb{P}_{\text{full}})$ denotes the marginal distribution of (X, Y, T) under $\mathbb{P}_{\text{full}}$.

We assume that we can factorize the full distribution as follows for rct and os

$$\mathbb{P}^{\diamond}_{\text{full}} = \underbrace{\mathbb{P}_{Y|Y(1),Y(0),T}}_{:=\mathbb{P}_{\text{det}}} \underbrace{\mathbb{P}_{Y(1),Y(0)|X,U}}_{:=\mathbb{P}_{\text{inv}}} \underbrace{\mathbb{P}^{\diamond}_{X,T,U}}_{:=\mathbb{P}^{\diamond}_{\text{cnf}}}, \tag{1}$$

where $\mathbb{P}_{\text{det}}$ is deterministically given by $Y = Y(T)$[2], $\mathbb{P}_{\text{inv}}$ is invariant across studies, and $\mathbb{P}^{\diamond}_{\text{cnf}}$ differs for $\diamond \in \{\text{rct}, \text{os}\}$. This factorization captures the essence of the potential outcome framework, where $Y(1)$ and $Y(0)$ do not depend on T while being more general. In particular, it allows for shifts in the marginal distribution of the observed and unobserved confounders.

We illustrate the corresponding graphical model in Figure 2. Note that numerous attempts have been made to unify potential outcomes and graphical models, with the most prominent being the Single World Intervention Graphs [50]. However, we propose a simpler graphical model in this context since we do not use the graph to infer counterfactual independencies.

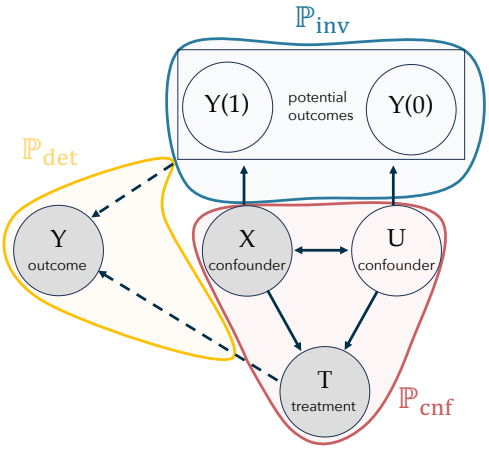


Figure 2: Graphical model that captures the Neyman-Rubin potential outcome framework with unobserved confounder U. $\mathbb{P}_{\text{inv}}$ is the causal mechanism that does not change between the randomized trial and the observational study, while $\mathbb{P}_{\text{cnf}}$ changes across studies. For the randomized trial, we assume there is no arrow from the confounders (X, U) to the treatment indicator T due to its internal validity. Observed variables are colored in shades of grey.

We now introduce three additional assumptions required for the validity of our statistical test and the resulting lower bound. First, we require transportability of the conditional average treatment effect (CATE).

Assumption 2.1 (Transportability)**.** *The conditional average treatment effect remains invariant across studies, that is*

$$\mathbb{E}_{\mathbb{P}^{\text{os}}_{\text{full}}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{\text{rct}}_{\text{full}}}[Y(1) - Y(0) \mid X].$$

This property is standard for generalizing the findings of randomized trials to another population [7, 13, 46], and is a weaker assumption than ignorability of study selection [28, 56] or sample ignorability of treatment effects [38].

Second, we assume that the randomized trial is internally valid.

Assumption 2.2 (Internal validity)**.** *The treatment is assigned independent of the covariates and the potential outcomes, that is,*

$$\mathbb{P}^{\text{rct}}_{\text{cnf}} = \mathbb{P}^{\text{rct}}_{T} \mathbb{P}^{\text{rct}}_{X,U}, \quad \text{with} \quad \mathbb{P}^{\text{rct}}_{T}(T = 1) = \pi \in (0, 1).$$

[2]Given that samples are drawn i.i.d., this asumption is equivalent to the classic SUTVA [51].

Internal validity holds by design in a completely randomized experiment, allowing for an unbiased estimation of the treatment effect. Observational studies, on the other hand, can have arbitrary confounding structures reflected in $\mathbb{P}_{\text{cnf}}$, i.e. $\mathbb{P}^{\text{os}}_{\text{cnf}} = \mathbb{P}^{\text{os}}_{T|X,U}\ \mathbb{P}^{\text{os}}_{X,U}$.

Finally, we assume that the population in the observational study includes the population in the trial.

Assumption 2.3 (Support inclusion)**.** *The support of the randomized trial is included in the support of the observational study, i.e.*

$$\text{supp}(\mathbb{P}^{\text{rct}}_X) \subseteq \text{supp}(\mathbb{P}^{\text{os}}_X).$$

This assumption is strictly weaker than the positivity of trial participation [2, 8, 24, 43, 54]. It is also expected to hold in our setting, as it aligns with the design of observational studies by regulatory agencies for drug monitoring [26], and particularly for post-marketing surveillance [20, 52].

2.1 Sensitivity analysis

Sensitivity analysis is commonly used to account for unobserved confounding in observational data. In particular, this approach estimates an interval for the treatment effect that depends on an assumed *confounding strength* Γ of $\mathbb{P}^{\text{os}}_{\text{cnf}}$. Throughout the paper, we define the confounding strength using the widely accepted marginal sensitivity model [57].

More formally, we assume that $\mathbb{P}^{\text{os}}_{\text{cnf}}$ belongs to the set $\mathcal{E}(\Gamma)$ of distributions that have bounded odds ratio,

$$\mathcal{E}(\Gamma) := \Big\{\mathbb{P}_{\text{cnf}} : \frac{1}{\Gamma} \leq \frac{\mathbb{P}_{\text{cnf}}(T=1 \mid X,U)}{\mathbb{P}_{\text{cnf}}(T=0 \mid X,U)} \Big/ \frac{\mathbb{P}_{\text{cnf}}(T=1 \mid X)}{\mathbb{P}_{\text{cnf}}(T=0 \mid X)} \leq \Gamma, \ \text{ a.s.}\Big\}.$$

Under this notion of confounding strength, we can define a set of full distributions $\tilde{\mathbb{P}}_{\text{full}}$ that are compatible with the marginal distribution of the observational study $\mathbb{P}^{\text{os}}$ and have a bounded odds ratio.

Definition 2.1 (Marginal sensitivity set)**.** *Given a distribution $\mathbb{P}^{\text{os}}$ over (X, Y, T) and a confounding strength $\Gamma \geq 1$, we define the set $\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma)$ of distributions $\tilde{\mathbb{P}}_{\text{full}}$, as*

$$\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma) := \{\tilde{\mathbb{P}}_{\text{full}} = \mathbb{P}_{\text{det}}\tilde{\mathbb{P}}_{\text{inv}}\tilde{\mathbb{P}}_{\text{cnf}} : \tilde{\mathbb{P}}_{\text{cnf}} \in \mathcal{E}(\Gamma) \quad \text{and} \quad \mathcal{M}(\mathbb{P}_{\text{det}}\tilde{\mathbb{P}}_{\text{inv}}\tilde{\mathbb{P}}_{\text{cnf}}) = \mathbb{P}^{\text{os}}\}.$$

In other words, this set contains all the full distributions that could have induced the marginal distribution of the observational study $\mathbb{P}^{\text{os}}$. Further, since the marginal sensitivity set contains $\mathbb{P}^{\text{os}}_{\text{full}}$ if Γ is well-specified, we can partially identify the (conditional) treatment effect as follows.

Definition 2.2 (Sensitivity bounds)**.** *We define the conditional average treatment effect (CATE) as*

$$\mu(X, \mathbb{P}_{\text{full}}) := \mathbb{E}_{\mathbb{P}_{\text{full}}}\left[Y(1) - Y(0) \mid X\right],$$

and the upper and lower bounds on CATE within the marginal sensitivity set as

$$\mu^+_\Gamma(X) := \sup_{\tilde{\mathbb{P}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma)} \mu(X, \tilde{\mathbb{P}}_{\text{full}}), \quad \mu^-_\Gamma(X) := \inf_{\tilde{\mathbb{P}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma)} \mu(X, \tilde{\mathbb{P}}_{\text{full}}).$$

Further, we define the average treatment effect (ATE) over a marginal distribution $\mathbb{P}_X$ that can differ from the marginal in $\mathbb{P}_{\text{full}}$ as

$$\mu(\mathbb{P}_X, \mathbb{P}_{\text{full}}) := \mathbb{E}_{\mathbb{P}_X}\left[\mu(X, \mathbb{P}_{\text{full}})\right],$$

and the upper and lower bounds on ATE as

$$\mu^+_\Gamma(\mathbb{P}_X) := \mathbb{E}_{\mathbb{P}_X}\left[\mu^+_\Gamma(X)\right], \quad \mu^-_\Gamma(\mathbb{P}_X) := \mathbb{E}_{\mathbb{P}_X}\left[\mu^-_\Gamma(X)\right].$$

Above, we do a slight abuse of notation by defining μ as both a function and a real number, depending on its argument. Several estimators have recently emerged in the literature for the CATE bounds [32, 35, 47] and for the ATE bounds [17, 18, 66]. We will leverage these estimators to construct a statistical test that detects unobserved confounding above a certain strength.

3 Methodology

We would like to test whether the unobserved full distribution $\mathbb{P}^{\text{os}}_{\text{full}}$, which marginalizes to $\mathbb{P}^{\text{os}}$, has confounding strength at most Γ. This is captured by the following null hypothesis

$$H_0(\Gamma) : \mathbb{P}^{\text{os}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma).$$

Note that in the special case where $\Gamma = 1$, the problem reduces to testing whether there are no unobserved confounders, i.e. $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$ under $\mathbb{P}^{\text{os}}$. We refer to this case, which has been recently studied in the literature (see Section 1.1), as *binary testing* for unobserved confounding.

In real-world scenarios, binary tests can be overly stringent, as they invalidate an observational study even if the unobserved confounding strength is negligible. To overcome this limitation, we propose the first test, to the best of our knowledge, for the general case where Γ is greater than one. In particular, underlying our testing procedure is a simple observation that follows from the sensitivity analysis bounds: When the null hypothesis is true for some confounding strength Γ, the average treatment effect under some target population should fall between the valid upper and lower bounds constructed from the observational study.

Lemma 3.1. *For any $\mathbb{P}_{\text{full}}$ which satisfies transportability, i.e. $\mu(X, \mathbb{P}_{\text{full}}) = \mu(X, \mathbb{P}^{\text{os}}_{\text{full}})$, and any $\mathbb{P}_X$ which satisfies support inclusion, i.e. $\text{supp}(\mathbb{P}_X) \subseteq \text{supp}(\mathbb{P}^{\text{os}}_X)$, it holds that*

$$\mathbb{P}^{\text{os}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma) \implies \mu(\mathbb{P}_X, \mathbb{P}_{\text{full}}) \in [\mu^-_\Gamma(\mathbb{P}_X), \mu^+_\Gamma(\mathbb{P}_X)].$$

Proof. First, note how $\mu(X, \mathbb{P}_{\text{full}}) \in [\mu^-_\Gamma(X), \mu^+_\Gamma(X)]$ for all $X \in \text{supp}(\mathbb{P}^{\text{os}}_X)$ when the null hypothesis $H_0(\Gamma)$ is true, due to the transportability assumption and the definition of CATE sensitivity bounds. The result then follows by taking expectations with respect to the corresponding marginals $\mathbb{P}_X$ on both sides. □

3.1 Statistical tests for $H_0(\Gamma)$

In what follows, we have access to a randomized trial $D_{\text{rct}} = \{(X_i, Y_i, T_i)\}_{i=1}^{n_{\text{rct}}}$ sampled i.i.d from the distribution $\mathbb{P}^{\text{rct}}$, and an observational study $D_{\text{os}} = \{(X_i, Y_i, T_i)\}_{i=1}^{n_{\text{os}}}$, sampled i.i.d. from the distribution $\mathbb{P}^{\text{os}}$. We first propose estimates for the average treatment effect under two target populations. Then, we leverage these estimates together with the sensitivity bounds to design an asymptotically valid statistical test at significance level α. Finally, we show how such a test can be used to establish an asymptotically valid lower bound on the unobserved confounding strength.

Estimating the ATE We discuss here how the average treatment effect can be estimated using data from the randomized trial. First, we define a target population $\mathbb{P}^{\diamond}_X$ to estimate the ATE. Then, the following lemma shows how the choice of $\mathbb{P}^{\text{rct}}_X$ and $\mathbb{P}^{\widetilde{\text{os}}}_X := \mathbb{P}^{\text{os}}_X \mid X \in \text{supp}(\mathbb{P}^{\text{rct}})$ allows us to identify ATE using data sampled from the randomized trial marginal distribution $\mathbb{P}^{\text{rct}}$.

Lemma 3.2. *For $\diamond \in \{\text{rct}, \widetilde{\text{os}}\}$, under Assumptions 2.1, 2.2 and 2.3, we have*

$$\mu(\mathbb{P}_X^{\diamond}, \mathbb{P}_{\text{full}}^{\diamond}) = \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[Y\left(\frac{T}{\pi} - \frac{(1-T)}{1-\pi}\right) w(X)\right], \quad \text{where} \quad w(X) := \frac{\mathbb{P}^{\diamond}(X)}{\mathbb{P}^{\text{rct}}(X)}.$$

Lemma 3.2 is a well-known result in the transportability literature [6, 9]. Essentially, it establishes that when the distribution shift between $\mathbb{P}_X^{\text{rct}}$ and $\mathbb{P}_X^{\diamond}$ can be corrected, we can identify and estimate the ATE under $\mathbb{P}_X^{\diamond}$.

Estimating the sensitivity interval Next, we discuss how $\mu_\Gamma^-(\mathbb{P}_X^{\diamond})$ and $\mu_\Gamma^+(\mathbb{P}_X^{\diamond})$ can be estimated using data from both the observational study and the target population $\mathbb{P}_X^{\diamond}$. Here, the approach varies based on the target population.

- For $\mathbb{P}_X^{\diamond} = \mathbb{P}_X^{\text{rct}}$, we estimate the CATE sensitivity bounds from observational data and average them over the target population. Specifically, we use the B-Learner [47] to estimate the sensitivity analysis bounds.
- For $\mathbb{P}_X^{\diamond} = \mathbb{P}_X^{\widetilde{\text{os}}}$, we have two options: either estimate the CATE sensitivity analysis bounds and average them, or directly estimate the ATE sensitivity analysis bounds over the target population. In our experiments, we directly estimate the ATE sensitivity analysis bounds using either the DVDS [18] or the QB estimator [17].

These methods yield estimates that are valid, sharp, and efficient under more general conditions than other existing methods. Nevertheless, our testing procedure is agnostic to the choice of the sensitivity analysis bound estimator, allowing for various options to be adopted.

Two statistical tests We outline our testing procedure in Algorithm 1, which can be instantiated for the target populations rct and $\widetilde{\text{os}}$. This results in two statistical tests, $\hat{\phi}_{\text{rct}}$ and $\hat{\phi}_{\widetilde{\text{os}}}$, for the null hypothesis $H_0(\Gamma)$. The following proposition confirms their asymptotic validity.

Proposition 3.1 (Validity of the test)**.** *Let $\hat{\phi}_\diamond(\Gamma, \alpha)$ be the test defined in Algorithm 1, for a fixed $\Gamma \in [1, \infty)$ and significance level α. Then, under Assumptions 2.1–2.3 and the setting described in Section 2, we have, for $H_0(\Gamma)$,*

(i) If it holds that $\lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} n_{\text{rct}}/n_{\text{os}} = 0$ and the estimators of the CATE sensitivity analysis bounds satisfy

$$\|\mu_\Gamma^{\pm} - \hat{\mu}_\Gamma^{\pm}\|_{L^2(\mathbb{P}^{\text{rct}})} = O_{\mathbb{P}^{\text{os}}}(n_{\text{os}}^{-1/2}),$$

$\hat{\phi}_{\text{rct}}(\Gamma, \alpha)$ is a valid asymptotic test at level α.

(ii) If $\hat{\mu}_\Gamma^+$ and $\hat{\mu}_\Gamma^-$ are consistent estimators of the ATE sensitivity analysis bounds that satisfy

$$\sqrt{n_{\text{os}}}\,\hat{\mu}_\Gamma^+ \xrightarrow{\mathcal{D}} \mathcal{N}(\mu_\Gamma^+, (\sigma_\Gamma^+)^2), \ \sqrt{n_{\text{os}}}\,\hat{\mu}_\Gamma^- \xrightarrow{\mathcal{D}} \mathcal{N}(\mu_\Gamma^-, (\sigma_\Gamma^-)^2),$$

$\hat{\phi}_{\widetilde{\text{os}}}(\Gamma, \alpha)$ is a valid asymptotic test at level α.

We provide a complete proof in Appendix A.1.2. Notably, Assumption (ii) is relatively mild and expected to hold for various estimators; for instance, it can be satisfied by the DVDS estimator [18]. On the other hand, Assumption (i) is stronger and generally only expected to hold when $n_{\text{os}} \gg n_{\text{rct}}$.

Algorithm 1 Statistical test for detecting unobserved confounding

1: **Input:** $\diamond \in \{\text{rct}, \widetilde{\text{os}}\}$, $D_{\text{rct}}, D_{\text{os}}$, significance level α, confounding strength Γ.

2: Estimate $\mu(\mathbb{P}^{\diamond}, \mathbb{P}^{\diamond}_{\text{full}})$ using the randomized trial dataset:

$$\hat{\mu} = \frac{1}{n_{\text{rct}}} \sum_{(X_i,T_i,Y_i)\in D_{\text{rct}}} Y_i \left(\frac{T_i}{\pi} - \frac{1-T_i}{1-\pi}\right) w(X_i), \quad \hat{\sigma}^2 = \widehat{\text{Var}}_{\mathbb{P}^{\text{rct}}}[\hat{\mu}].$$

3: Estimate the sensitivity analysis bounds $\hat{\mu}^-_\Gamma(X)$ and $\hat{\mu}^+_\Gamma(X)$ using the observational study dataset, and average over the target population $\mathbb{P}^{\diamond}$:

$$\hat{\mu}^+_\Gamma = \hat{\mathbb{E}}_{\mathbb{P}^{\diamond}_X}[\hat{\mu}^+_\Gamma(X)], \quad (\hat{\sigma}^+_\Gamma)^2 = \widehat{\text{Var}}_{\mathbb{P}^{\diamond}_X}[\hat{\mu}^+_\Gamma(X)], \quad \hat{\mu}^-_\Gamma = \hat{\mathbb{E}}_{\mathbb{P}^{\diamond}_X}[\hat{\mu}^-_\Gamma(X)], \quad (\hat{\sigma}^-_\Gamma)^2 = \widehat{\text{Var}}_{\mathbb{P}^{\diamond}_X}[\hat{\mu}^-_\Gamma(X)],$$

where $\hat{\mathbb{E}}[\cdot]$ and $\widehat{\text{Var}}[\cdot]$ denote the empirical mean and variance, respectively.

4: Compute the test statistics:

$$\hat{T}^+_\Gamma = \frac{\hat{\mu}^+_\Gamma - \hat{\mu}}{\hat{\sigma}^+_\diamond}, \quad \text{where} \quad \hat{\sigma}^+_{\text{rct}} = \sqrt{(\hat{\sigma}^+_\Gamma)^2 + \hat{\sigma}^2 + 2\hat{\sigma}^+_\Gamma\hat{\sigma}} \quad \text{and} \quad \hat{\sigma}^+_{\widetilde{\text{os}}} = \sqrt{(\hat{\sigma}^+_\Gamma)^2 + \hat{\sigma}^2},$$

$$\hat{T}^-_\Gamma = \frac{\hat{\mu} - \hat{\mu}^-_\Gamma}{\hat{\sigma}^-_\diamond}, \quad \text{where} \quad \hat{\sigma}^-_{\text{rct}} = \sqrt{(\hat{\sigma}^-_\Gamma)^2 + \hat{\sigma}^2 + 2\hat{\sigma}^-_\Gamma\hat{\sigma}} \quad \text{and} \quad \hat{\sigma}^-_{\widetilde{\text{os}}} = \sqrt{(\hat{\sigma}^-_\Gamma)^2 + \hat{\sigma}^2}.$$

5: **Output:** $\hat{\phi}_\diamond(\Gamma, \alpha) = \mathbb{I}\{\min\left(\hat{T}^+_\Gamma, \hat{T}^-_\Gamma\right) < z_{\alpha/2}\}$, where z_α is the α-quantile of the standard normal.

In essence, we propose two tests that work under different assumptions: $\hat{\phi}_{\text{rct}}$ relies on a consistent estimate of the CATE sensitivity analysis bounds, while $\hat{\phi}_{\widetilde{\text{os}}}$ requires an estimate of the importance weights $w(X)$[3].

Advantages of each test The test $\hat{\phi}_{\widetilde{\text{os}}}$ can be advantageous when CATE estimation is challenging (e.g. when the outcomes are binary and the classes are imbalanced or when the observational study has a limited sample size), but the weights $w(X)$ can be identified, and vice versa for the test $\hat{\phi}_{\text{rct}}$. In addition, $\hat{\phi}_{\widetilde{\text{os}}}$ can benefit from large observational studies as the variances $(\hat{\sigma}^-_\Gamma)^2$ and $(\hat{\sigma}^+_\Gamma)^2$ vanish for large n_{os}.

3.2 A lower bound on unobserved confounding strength

The statistical test described in the previous section raises a question about what level of confounding strength is reasonable to test. Ideally, epidemiologists would like to estimate the confounding strength instead of conducting a test. However, this is infeasible unless the support of $\mathbb{P}^{\text{rct}}_X$ and $\mathbb{P}^{\text{os}}_X$ are the same.

A practical alternative is to estimate a lower bound on the true unobserved confounding strength defined as

$$\Gamma^\star := \inf\{\Gamma : \mathbb{P}^{\text{os}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma)\}.$$

Given an observational study and a randomized trial, we aim to find a quantity that, with high probability, is a lower bound for the true confounding strength $\Gamma^\star$. Without loss of generality, we fix the test $\hat{\phi}_{\text{rct}}$ and recall that $\hat{\phi}_{\text{rct}}(\Gamma, \alpha)$ is a deterministic function given the data[4]. Hence, we obtain a lower bound for a fixed significance level α by computing

$$\hat{\Gamma}_{\texttt{LB}} = \inf_\Gamma\{\Gamma : \hat{\phi}_{\text{rct}}(\Gamma, \alpha) = 0\}, \tag{2}$$

[3]The importance weights can be identified when the observational study and the randomized trial adhere to a nested trial design [5, 44, 45]. See Appendix A.2 for a discussion on how the importance weights are estimated in this setting.

[4]When bootstrap is used to estimate the variance we fix the bootstrap bags for all Γ.

that is, in words, the smallest Γ such that the test accepts the null hypothesis. In practice, we compute $\hat{\Gamma}_{\texttt{LB}}$ with a grid search over values of Γ starting from 1 until the first test acceptance.

We show in the following proposition that $\hat{\Gamma}_{\texttt{LB}}$ is a valid lower bound for $\Gamma^\star$.

Proposition 3.2. *Let $\hat{\Gamma}_{LB}$ be as in Equation (2) for a fixed significance level α. Then, under Assumptions 2.1–2.3 and the setting described in Section 2, $\hat{\Gamma}_{LB}$ is an asymptotically valid lower bound, i.e.*

$$\mathbb{P}(\hat{\Gamma}_{LB} \leq \Gamma^\star) \geq 1 - \alpha - o_{\mathbb{P}}(1).$$

Proof. Note that by definition of $\hat{\Gamma}_{\texttt{LB}}$, we have that

$$\begin{aligned}\mathbb{P}(\hat{\Gamma}_{\texttt{LB}} > \Gamma^\star) &= \mathbb{P}(\cap_{\Gamma \leq \Gamma^\star}\{\hat{\phi}_{\text{rct}}(\Gamma, \alpha) = 1\}) \\ &\leq \mathbb{P}(\hat{\phi}_{\text{rct}}(\Gamma^\star, \alpha) = 1) \leq \alpha + o_{\mathbb{P}}(1),\end{aligned}$$

where the last inequality follows from the asymptotic validity of the test in Proposition 3.1. □

4 Synthetic Experiments

In this section, we evaluate our two tests and the resulting lower bounds in finite-sample synthetic and semi-synthetic experiments. In particular, we fix the true unobserved confounding strength $\Gamma^\star$ and conduct experiments varying the sample size and the invariant distribution $\mathbb{P}_{\text{inv}}$.

First, we postulate that, for a fixed $\Gamma^\star$, the tightness of the lower bound $\hat{\Gamma}_{\texttt{LB}}$ improves when the confounder U is more informative about the potential outcomes $(Y(1), Y(0))$. In our experiments, we choose the correlation between the unobserved confounder and one of the potential outcomes as a proxy measure of information,

$$\rho_{u,y} = \frac{\text{Cov}_{\mathbb{P}^{\text{os}}_{\text{full}}}[Y(1), U]}{\sigma_{Y(1)}\sigma_U}. \tag{3}$$

Intuitively, the sensitivity analysis bounds are tight for a specific $\Gamma^\star$ when $\mathbb{P}^{\text{os}}_{\text{cnf}}$ leads to a marginal distribution $\mathbb{P}^{\text{os}}$ that maximally biases the estimable ATE. This situation occurs, for instance, when patients experiencing smaller outcomes are assigned to the control group while those with larger outcomes are in the treatment group. In this case, the sensitivity analysis bounds must be sufficiently large to include the true ATE and remain valid. Such a scenario is only possible if U is very informative of $Y(1)$, captured by a high correlation coefficient. Conversely, when $Y(1)$ is independent of U, the true ATE is unaffected by the unobserved confounding, and the sensitivity bounds are unnecessarily conservative, leading to low power of the test and hence looser $\hat{\Gamma}_{\texttt{LB}}$. More formally, in Appendix A.3 we show that when the confounder U is equal to $(Y(1), Y(0))$, the correlation coefficient $\rho_{u,y} = 1$ and $\hat{\Gamma}_{\texttt{LB}}$ converges to $\Gamma^\star$ in the infinite sample limit. In contrast, when U is independent of $(Y(1), Y(0))$, $\rho_{u,y} = 0$ and $\hat{\Gamma}_{\texttt{LB}} = 1$.

Second, we study the behavior of the lower bound as the observational study sample size grows. In real-world situations, increasing the number of samples in a randomized trial is often constrained by the logistical challenges of conducting additional experiments. However, observational studies have the potential for continuous growth through electronic health records and insurance claims databases. In the context of postmarketing, ongoing monitoring enables the inclusion of data from newly exposed individuals. Therefore, we compare our two tests when the sample size of the observational study grows: our experiments show that $\hat{\phi}_{\widetilde{\text{os}}}$ has better statistical power when n_{os} is large.

4.1 Datasets

Synthetic distribution We first benchmark tests and respective lower bounds with a synthetic distribution similar to [33, 64]. Here, the propensity score, the true unobserved confounding strength $\Gamma^\star$, and the correlation strength can be designed.

We choose the invariant $\mathbb{P}_{\text{inv}}$ to be the following linear outcome model

$$Y(T) = (2T-1)X + (2T-1) + U + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_Y^2).$$

For the marginal distribution over X, U in $\mathbb{P}^{\diamond}_{\text{cnf}}$ we generate an unobserved confounder $U \sim \text{Unif}(0,1)$ for both study designs and draw the observed covariate according to

$$\mathbb{P}_X^{\text{rct}} = \text{Unif}(-1,1),\ \mathbb{P}_X^{\text{os}} = \text{Unif}(-2,2).$$

Further, for the observational distribution, we choose the conditional distribution of the treatment T given X, U to be a Bernoulli, which satisfies the marginal sensitivity model with an odds ratio equal to $\Gamma^\star$. Specifically, we fix the marginal propensity score as

$$\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid X) = \text{logistic}\left(0.75X + 0.5\right),$$

and design the full propensity score $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid X, U)$ such that it marginalizes to $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid X)$. For the randomized control trial, we choose $\pi = \mathbb{P}^{\text{rct}}_{\text{cnf}}(T = 1|X, U) = 1/2$. We refer the reader to Appendix B.1 for complete experimental details.

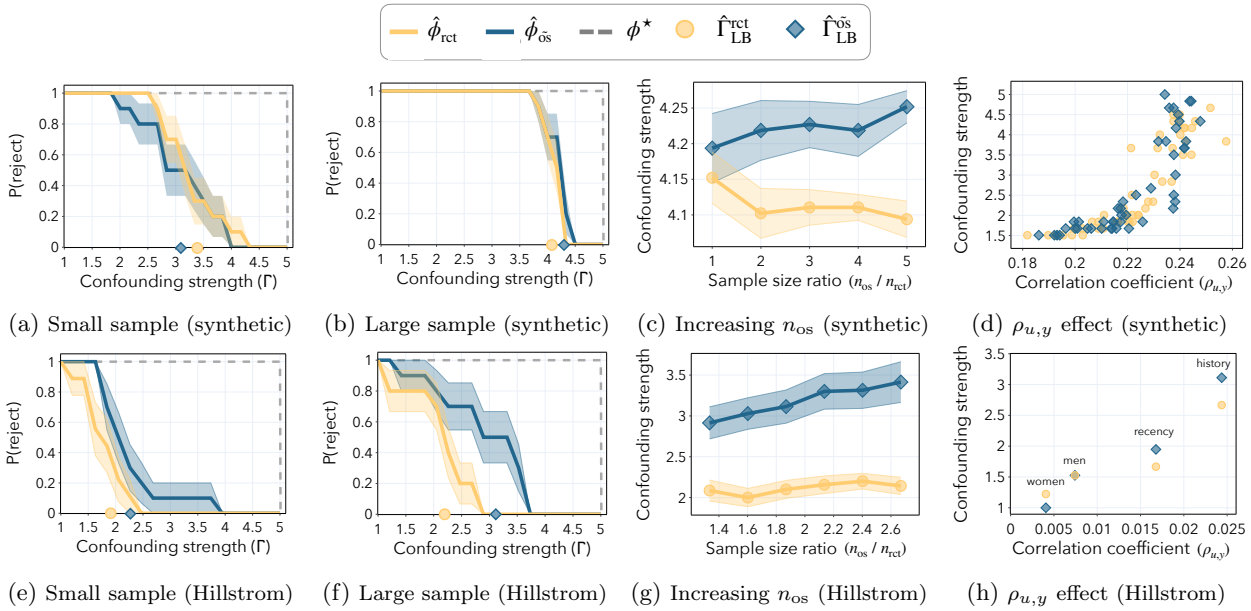
Semi-synthetic datasets We expand our benchmark using three real-world randomized trials: Hillstrom's MineThatData Email data [27], the Tennessee STAR study [63] and the VOTE dataset [22]. In contrast to the synthetic experiments, these datasets involve real outcome functions, though the treatment assignment is still controlled.

We focus on Hillstrom's dataset for clarity of presentation, and we refer the reader to Appendix C.2 for experiments on the other datasets showing similar trends. Hillstrom [27] focused on measuring the impact of an email campaign on the dollars spent by the recipients in the following two weeks. We first sample a small subset of the original trial, D, as our randomized trial, D_{rct}. We can then subsample multiple observational studies from $D \setminus D_{\text{rct}}$ sharing a fixed true confounding strength $\Gamma^\star$, i.e. $\mathbb{P}_{\text{cnf}}$, but with a varying correlation between the hidden confounder U and outcome $Y(1)$, i.e. $\mathbb{P}_{\text{inv}}$.

Let us denote X_{all} as the vector of all observed covariates. While we cannot intervene on $\mathbb{P}_{\text{inv}}(Y(1), Y(0)|X_{\text{all}})$ as it is intrinsic to the dataset, we can generate multiple observational studies by partitioning X_{all} into unobserved U and observed X in different ways. For a given partitioning $X_{\text{all}} = (U, X)$, the resulting D_{os} will have a specific $\mathbb{P}_{\text{inv}}(Y(1), Y(0)|U)$ and hence correlation coefficient $\rho_{u,y}$. With each choice of U, we enforce a propensity score $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$ that satisfies $\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma^\star)$ by subsampling $D \setminus D_{\text{rct}}$. Finally, we remove U to construct D_{os}. Our subsampling approach is a variation of the methods presented in Gentzel et al. [21], Keith et al. [37] (see further details in Appendix B.2). Finally, we enforce Assumption 2.3 by excluding urban zip codes from the support of the randomized trial.

4.2 Experimental results

We now discuss our experimental results depicted in Figure 3. The top row presents results for the synthetic experiments, and the bottom row for the semi-synthetic experiments.



(a) Small sample (synthetic) (b) Large sample (synthetic) (c) Increasing n_{os} (synthetic) (d) $\rho_{u,y}$ effect (synthetic)

(e) Small sample (Hillstrom) (f) Large sample (Hillstrom) (g) Increasing n_{os} (Hillstrom) (h) $\rho_{u,y}$ effect (Hillstrom)

Figure 3: For all the plots: the significance level is $\alpha = 0.05$, $\phi^\star$ denotes the oracle test which rejects for $\Gamma < \Gamma^\star$, $\hat{\Gamma}^{\text{rct}}_{\text{LB}}$ and $\hat{\Gamma}^{\widetilde{\text{os}}}_{\text{LB}}$ denote which test is used to compute $\hat{\Gamma}_{\text{LB}}$. First row with *synthetic experiment* choosing $\Gamma^\star = 5$: Probability of rejection for different Γ and average $\hat{\Gamma}_{\text{LB}}$ for the test for (a) small sample size: $n_{\text{rct}} = 2K, n_{\text{os}} = 2K$ and (b) large sample size: $n_{\text{rct}} = 20K, n_{\text{os}} = 20K$. $\hat{\Gamma}_{\text{LB}}$ for (c) increasing sample size of the observational study with $n_{\text{rct}} = 20K$ and (d) increasing correlation coefficient; $n_{\text{rct}} = 20K, n_{\text{os}} = 20K$. Second row with the *semi-synthetic* Hillstrom dataset choosing $\Gamma^\star = 5$ and using "history" as unobserved confounder (except in (h)): Probability of rejection for different Γ and average $\hat{\Gamma}_{\text{LB}}$ for (e) small sample size: $n_{\text{rct}} = 2300$, $n_{\text{os}} = 6150$ and (f) large sample size: $n_{\text{rct}} = 7680$, $n_{\text{os}} = 20500$. $\hat{\Gamma}_{\text{LB}}$ for (g) increasing n_{os} with $n_{\text{rct}} = 7680$ and (h) increasing correlation coefficient.

Effect of observational study sample size First, we observe in Figures 3a-3b and Figures 3e-3f that our tests are valid in all settings, i.e. they do not reject for strengths larger than $\Gamma^\star$. However, the statistical power substantially improves in the large sample size regime. In general, the performance of both tests aligns. In Figures 3c and 3g, the lower bounds $\hat{\Gamma}^{\text{rct}}_{\text{LB}}$ and $\hat{\Gamma}^{\widetilde{\text{os}}}_{\text{LB}}$ vary with the sample size of the observational study. We confirm that the $\hat{\phi}_{\widetilde{\text{os}}}$ derives greater benefits from a larger observational study sample size than $\hat{\phi}_{\text{rct}}$, as discussed in Section 3.1.

Effect of outcome-confounder correlation Note that the tests in Figure 3a-3b and Figure 3e-3f are somewhat conservative: The probability of rejection for Γ close to $\Gamma^\star$ is small, which leads to a rather loose lower bound estimate $\hat{\Gamma}_{\text{LB}}$. This is due to a fundamental limitation of the marginal sensitivity model that cannot be overcome without additional assumptions on how U affects Y, as discussed in Appendix A.3. We study here the effect of increasing the outcome-confounder correlation (Equation 3). Specifically, we generate observational datasets with a constant $\Gamma^\star$ but varying $\rho_{u,y}$, and report $\hat{\Gamma}_{\text{LB}}$ for both tests. For the synthetic experiments in Figure 3d, we plot $\hat{\Gamma}^{\text{rct}}_{\text{LB}}$ and $\hat{\Gamma}^{\widetilde{\text{os}}}_{\text{LB}}$ for $n = 50$ distinct values of $\sigma^2_{Y(1)} \sim \text{Unif}[0, 1]$. For the semi-synthetic experiments in Figure 3h, we depict $\hat{\Gamma}^{\text{rct}}_{\text{LB}}$ and $\hat{\Gamma}^{\widetilde{\text{os}}}_{\text{LB}}$ for different hidden confounders U. Both plots confirm our hypothesis that higher $\rho_{u,y}$ correlates with a tighter lower bound $\hat{\Gamma}_{\text{LB}}$.

5 Real-world experiments

Linking back to the pipeline in Figure 1, we demonstrate how epidemiologists can use the lower bound $\hat{\Gamma}_{\texttt{LB}}$ to successfully differentiate between studies with significant confounding and those with negligible confounding. Specifically, we propose comparing $\hat{\Gamma}_{\texttt{LB}}$ with a critical value of Γ, estimated from the available observational data

$$\hat{\Gamma}_{\texttt{CT}} := \inf\{\Gamma : 0 \in [\hat{\mu}_{\Gamma}^{-}, \hat{\mu}_{\Gamma}^{+}]\}.$$

In essence, $\hat{\Gamma}_{\texttt{CT}}$ represents the minimum strength for which sensitivity analysis includes both positive and negative values of treatment effect, thereby invalidating the study results. Similar critical values have been proposed in the literature to assess the robustness of conclusions drawn from observational data, see e.g. Jin et al. [33], VanderWeele and Ding [59]. The most appropriate choice for the specific context should be determined by epidemiologists.

We flag an observational study as confounded if $\hat{\Gamma}_{\texttt{LB}}$ exceeds the critical value, i.e.

$$\psi_{\texttt{sensi}} := \mathbb{I}\{\hat{\Gamma}_{\texttt{LB}} > \hat{\Gamma}_{\texttt{CT}}\}. \tag{4}$$

We compare our decision-making procedure with one based on a binary test

$$\psi_{\texttt{bin}} = \mathbb{I}\{\hat{\Gamma}_{\texttt{LB}} > 1\}.$$

In contrast to our procedure, the output of the binary one flags an observational study if any level of confounding is detected. Note that choosing a more powerful binary test in the literature would only *exacerbate* this issue.

Controversy around HRT For years, epidemiologists could not reach a consensus on the impact of hormone replacement therapy (HRT) on coronary heart disease and stroke based on the findings of the Women's Health Initiative (WHI) study [1]. The WHI study included a randomized trial and an observational study that examined the impact of HRT on various cardiovascular events. While the observational study suggested that HRT had a protective effect against these outcomes, the randomized trial indicated the opposite. This discrepancy was recently resolved by identifying a strong unobserved confounder - the time t since the start of HRT - and reanalyzing the data accordingly [58]. We now present evidence that our procedure can yield the same epidemiological conclusions and avoid issuing false alarms when the confounding is negligible.

Experimental details We consider two binary-valued outcomes: the presence of stroke and coronary heart disease within the follow-up period. We apply our procedure from Equation (4) to both the original dataset, which includes all patients (i.e. $t \leq 20$), and a subsampled dataset that only includes patients who were not previous users of HRT (i.e. $t = 0$). Since the WHI study satisfies the criteria for a nested trial design, we calculate $\hat{\Gamma}_{\texttt{LB}}$ using our testing procedure $\hat{\phi}_{\widetilde{\text{os}}}$. See Appendix B.3 for experimental details.

Results In Table 1, we show the result of both procedures on the WHI dataset, with small ($t = 0$) and large ($t \leq 20$) unobserved confounding .

For coronary heart disease, both algorithms flag the study as confounded when strong unobserved confounding is present ($t \leq 20$). However, when minimal unobserved confounding is present ($t = 0$), our test does not flag the study, while $\psi_{\texttt{bin}}$ does. This difference underscores our test's capability to distinguish between

Table 1: The significance level is $\alpha = 0.05$. For $t = 0$ (small confounding), the study only included patients who were not previous users of HRT. For $t \leq 20$ (strong confounding), the study includes patients who have been using HRT for up to 20 years.

Metric	Stroke		Coronary heart disease	
	$t = 0$	$t \leq 20$	$t = 0$	$t \leq 20$
$\hat{\Gamma}_{\texttt{CT}}$	1.017	1.172	1.017	1.164
$\hat{\Gamma}_{\texttt{LB}}$	1.052	1.207	1.009	1.224
$\psi_{\texttt{bin}}$	1	1	1	1
$\psi_{\texttt{sensi}}$	1	1	0	1

small and large unobserved confounding, thereby addressing a limitation in the flagging procedures based on existing testing methods.

In the case of stroke, both $\psi_{\texttt{sensi}}$ and $\psi_{\texttt{bin}}$ correctly flag the observational study, even when we adjust for the time since the start of treatment ($t = 0$). This finding aligns with experts suggesting that additional unobserved confounding factors for stroke are still present after controlling for the time since the start of hormone replacement therapy [49].

Observe that an alternative way to reach the same conclusions is by testing the difference in ATE estimates between the two studies. However, our approach offers a notable advantage: it allows us to test if the observational study is too confounded on arbitrarily fine-grained subgroups up to the individual level. Indeed, we can estimate the CATE sensitivity analysis bounds and compare critical values for specific subgroups against our lower bound. In contrast, testing differences in group-level ATE estimates would require several tests, one for each subgroup, leading to issues with multiple testing and insufficient sample sizes.

6 Discussion and future work

Our approach shares limitations with other methods that test for unobserved confounding. Since we rely on the transportability assumption, our test could misidentify violations of this assumption as unobserved confounding. In addition, the lower bound we provide is optimistic; outside the common support of the two studies, the unobserved confounding could be arbitrarily high. Furthermore, our test is designed to detect confounding structures that bias the average treatment effect and, hence, would not detect confounding bias that cancels out on average.

Our discussion suggests several important directions for future research. First, developing a more refined sensitivity model that accounts for the correlation between outcomes and unobserved confounders could result in a more powerful test. Second, our test could be adapted to the scenario where multiple observational datasets may be available but no randomized control trials. Lastly, it would be highly valuable to propose a procedure that not only identifies hidden confounding but also suggests specific interventions to mitigate it.

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

[2] Isaiah Andrews and Emily Oster. Weighting for external validity. *NBER: National Bureau of Economic Research*, 2017.

[3] Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *arXiv preprint arXiv:2302.13428*, 2023.

[4] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[5] Niteesh Choudhry. Randomized, controlled trials in health insurance systems. *New England Journal of Medicine*, 377(10):957–964, 2017.

[6] Stephen Cole and Elizabeth Stuart. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115, 2010.

[7] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

[8] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Reweighting the RCT for generalization: finite sample analysis and variable selection. *arXiv preprint arXiv:2208.07614*, 2022.

[9] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023.

[10] Jerome Cornfield, William Haenszel, Cuyler Hammond, Abraham Lilienfeld, Michael Shimkin, and Ernst Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *JNCI: Journal of the National Cancer Institute*, 22(1):173–203, 01 1959.

[11] Issa Dahabreh, Sebastien Haneuse, James Robins, Sarah Robertson, Ashley Buchanan, Elizabeth Stuart, and Miguel Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190(8):1632–1642, 2021.

[12] Xavier De Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2(2):187–199, 2014.

[13] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

[14] Irina Degtiar, Tim Layton, Jacob Wallace, and Sherri Rose. Conditional cross-design synthesis estimators for generalizability in Medicaid. *Biometrics*, 00:1 – 14, 2023.

[15] 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.

[16] Stephen Donald, Yu-Chin Hsu, and Robert Lieli. Testing the unconfoundedness assumption via inverse probability weighted estimators of (L) ATT. *Journal of Business & Economic Statistics*, 32(3):395–415, 2014.

[17] Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, pages 1–13, 2022.

[18] Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

[19] Nancy Dreyer. Advancing a framework for regulatory use of real-world evidence: when real is reliable. *Therapeutic Innovation and Regulatory Science*, 52(3):362–368, 2018.

[20] Jessica Franklin, Robert 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.

[21] Amanda Gentzel, Purva Pruthi, and David Jensen. How and why to use experimental data to evaluate methods for observational causal inference. In *International Conference on Machine Learning*, 2021.

[22] Alan S Gerber, Donald P Green, and Christopher W Larimer. Social pressure and voter turnout: evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48, 2008.

[23] Laura Goetz and Nicholas Schork. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963, 2018.

[24] 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: Statistics in Society*, 178(3):757–778, 2015.

[25] Erika Check Hayden. Gene therapies pose million-dollar conundrum. *Nature*, 534:305–306, 2016.

[26] Zhe He, Xiang Tang, Xi Yang, Yi Guo, Thomas 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.

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

[28] Joseph Hotz, Guido Imbens, and Julie Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of econometrics*, 125(1-2):241–270, 2005.

[29] Jesse Hsu and Dylan Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811, 2013.

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

[31] 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.

[32] 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.

[33] Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

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

[35] 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.

[36] Rickard Karlsson and Jesse Krijthe. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, 2023.

[37] Katherine Keith, Sergey Feldman, David Jurgens, Jonathan Bragg, and Rohit Bhattacharya. RCT rejection sampling for causal estimation evaluation. *arXiv preprint arXiv:2307.15176*, 2023.

[38] Holger Kern, Elizabeth Stuart, Jennifer Hill, and Donald Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127, 2016.

[39] David 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.

[40] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383, 2010.

[41] Nicolai Meinshausen. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

[42] 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.

[43] Xinkun Nie, Guido Imbens, and Stefan Wager. Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*, 2021.

[44] Manfred Olschewski and Martin Scheurlen. Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. *Methods of Information in Medicine*, 24(03):131–134, 1985.

[45] Manfred Olschewski, Martin Schumacher, and Kathryn B Davis. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled Clinical Trials*, 13(3):226–239, 1992.

[46] Colm O'Muircheartaigh and Larry Hedges. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2): 195–210, 2014.

[47] 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.

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

[49] Ross Prentice, Robert Langer, Marcia Stefanick, Barbara Howard, Mary Pettinger, Garnet Anderson, David Barad, 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.

[50] Thomas Richardson and James Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[51] Donald Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

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

[53] Tamar Sofer, David Richardson, Elena Colicino, Joel Schwartz, and Eric Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

[54] Elizabeth Stuart, Stephen Cole, Catherine Bradshaw, and Philip Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

[55] Alexander Stutz. Can semi-supervised learning improve the estimation of causal treatment effects? Master's thesis, ETH Zurich, August 2023.

[56] RA Sugden and TMF Smith. Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495–506, 1984.

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

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

[59] Tyler VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

[60] Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 2020.

[61] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph 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.

[62] Vera Vlahović-Palčevski and Dirk Mentzer. Postmarketing surveillance. *Pediatric Clinical Pharmacology*, pages 339–351, 2011.

[63] Elizabeth Word, John Johnston, Helen Bain, DeWayne Fulton, Jayne Zaharias, Charles Achilles, Martha Lintz, John Folger, and Carolyn Breda. The state of Tennessee's student/teacher achievement ratio (STAR) project. *Tennessee Board of Education*, 1990.

[64] Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50 (5):2587–2615, 2022.

[65] 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.

[66] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.

Appendices

The following appendices provide deferred proofs, experiment details, and ablation studies.

Table of contents

A Methodology

A.1 Remaining proofs

We present here the proofs for Lemma 3.2 and Proposition 3.1.

A.1.1 Proof of Lemma 3.2

For $\diamond = \text{rct}$, we have

$$\begin{aligned}\mu(\mathbb{P}_X^{\text{rct}}, \mathbb{P}_{\text{full}}^{\text{rct}}) &= \mathbb{E}_{\mathbb{P}_{\text{full}}^{\text{rct}}}[Y(1) - Y(0)] \\ &= \mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\left[\mathbb{E}_{\mathbb{P}^{\text{rct}}}[Y \mid T=1, X]\frac{\mathbb{P}^{\text{rct}}(T=1)}{\mathbb{P}^{\text{rct}}(T=1)} - \mathbb{E}_{\mathbb{P}^{\text{rct}}}[Y \mid T=0, X]\frac{\mathbb{P}^{\text{rct}}(T=0)}{\mathbb{P}^{\text{rct}}(T=0)}\right] \\ &= \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[Y\left(\frac{T}{\pi} - \frac{(1-T)}{1-\pi}\right)\right],\end{aligned}$$

where the last equality follows from the internal validity of the randomized trial.

For $\diamond = \widetilde{\text{os}}$, we first note that by transportability of CATE and definition of $\mathbb{P}^{\widetilde{\text{os}}}$, we have

$$\mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\widetilde{\text{os}}}) = \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}}) = \mathbb{E}_{\mathbb{P}_X^{\widetilde{\text{os}}}}\left[\mathbb{E}_{\mathbb{P}_{\text{full}}^{\text{os}}}[Y(1) - Y(0) \mid X]\right] = \mathbb{E}_{\mathbb{P}_X^{\widetilde{\text{os}}}}\left[\mathbb{E}_{\mathbb{P}_{\text{full}}^{\text{rct}}}[Y(1) - Y(0) \mid X]\right].$$

Furthermore, it holds via Assumption 2.3 (support inclusion) and the definition of $\mathbb{P}^{\widetilde{\text{os}}}$ that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\left[\mathbb{E}_{\mathbb{P}_{\text{full}}^{\text{rct}}}[Y(1) - Y(0) \mid X]\frac{\mathbb{P}^{\widetilde{\text{os}}}(X)}{\mathbb{P}^{\text{rct}}(X)}\right] &= \mathbb{E}_{\mathbb{P}_X^{\text{rct}}}\left[\left(\mathbb{E}_{\mathbb{P}^{\text{rct}}}[Y \mid T=1, X] - \mathbb{E}_{\mathbb{P}^{\text{rct}}}[Y \mid T=0, X]\right)\frac{\mathbb{P}^{\widetilde{\text{os}}}(X)}{\mathbb{P}^{\text{rct}}(X)}\right] \\ &= \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[Y\left(\frac{T}{\pi} - \frac{(1-T)}{1-\pi}\right)\frac{\mathbb{P}^{\widetilde{\text{os}}}(X)}{\mathbb{P}^{\text{rct}}(X)}\right],\end{aligned}$$

where the last equality again follows from the internal validity of the randomized trial.

A.1.2 Proof of Proposition 3.1

First, observe that by definition,

$$\{\hat{\phi}_\diamond(\Gamma, \alpha) = 1\} \implies \{\hat{T}_\Gamma^+ \le z_{\alpha/2}\} \cup \{\hat{T}_\Gamma^- \le z_{\alpha/2}\}. \tag{5}$$

Hence, if $\mathbb{P}_{H_0(\Gamma)}(\hat{T}_\Gamma^- \le z_{\alpha/2}) \le \frac{\alpha}{2} + o_{\mathbb{P}}(1)$ and $\mathbb{P}_{H_0(\Gamma)}(\hat{T}_\Gamma^+ \le z_{\alpha/2}) \le \frac{\alpha}{2} + o_{\mathbb{P}}(1)$, the theorem follows from the union bound. For brevity, we only prove $\hat{T}_\Gamma^+ \le z_{\alpha/2}$ as the proof for $\hat{T}_\Gamma^-$ is analogous.

Proof of case $\diamond = \text{rct}$ Let (X_i, Y_i, T_i) be i.i.d. sampled from $\mathbb{P}^{\text{rct}}$ and define

$$\begin{aligned}&Z = \left(\frac{YT}{\pi} - \frac{Y(1-T)}{1-\pi}, \quad \mu_\Gamma^+(X)\right)^T, \quad \text{with} \\ &\mu := \mathbb{E}_{\mathbb{P}^{\text{rct}}}[Z] = \left(\mu\left(\mathbb{P}_X^{\text{rct}}, \mathbb{P}_{\text{full}}^{\text{rct}}\right), \quad \mu_\Gamma^+\left(\mathbb{P}_X^{\text{rct}}\right)\right)^T \quad \text{and} \quad \Sigma := \text{Cov}_{\mathbb{P}^{\text{rct}}}(Z) = \begin{pmatrix} \sigma^2 & \xi_\Gamma \\ \xi_\Gamma & (\sigma_\Gamma^+)^2 \end{pmatrix} < \infty.\end{aligned}$$

Further, we define

$$\bar{Z}_n = \left(\frac{1}{n_{\text{rct}}} \sum_{(T_i,Y_i)\in D_{\text{rct}}} \frac{Y_i T_i}{\pi} - \frac{Y_i(1-T_i)}{1-\pi}, \frac{1}{n_{\text{rct}}} \sum_{X_i \in D_{\text{rct}}} \mu_\Gamma^+(X_i) \right).$$

By the multivariate central limit theorem, we have

$$\sqrt{n_{\text{rct}}} \left(\bar{Z}_n - \mu \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \quad \text{as} \quad n_{\text{rct}} \to \infty.$$

Thus, it follows from the Cramér-Wold theorem that

$$\sqrt{n_{\text{rct}}} \left(\frac{1}{n_{\text{rct}}} \sum_{X_i \in D_{\text{rct}}} \mu_\Gamma^+(X_i) - \hat{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu_\Gamma^+\left(\mathbb{P}_X^{\text{rct}}\right) - \mu\left(\mathbb{P}_X^{\text{rct}}, \mathbb{P}_{\text{full}}^{\text{rct}}\right), \sigma^2 + (\sigma_\Gamma^+)^2 - 2\xi_\Gamma \right), \quad \text{as} \quad n_{\text{rct}} \to \infty.$$

It remains to show that asymptotic normality also holds when we use the empirical estimate $\hat{\mu}_\Gamma^+$. To do so, we prove the following convergence in probability

$$\left| \frac{1}{\sqrt{n_{\text{rct}}}} \sum_{X_i \in D_{\text{rct}}} \mu_\Gamma^+(X_i) - \hat{\mu}_\Gamma^+(X_i) \right| = o_{\mathbb{P}^{\text{os}}}(1), \quad \text{as} \quad n_{\text{rct}} \to \infty. \tag{6}$$

Then, by Slutsky's theorem, we have

$$\sqrt{n_{\text{rct}}} \left(\hat{\mu}_\Gamma^+ - \hat{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu_\Gamma^+\left(\mathbb{P}_X^{\text{rct}}\right) - \mu\left(\mathbb{P}_X^{\text{rct}}, \mathbb{P}_{\text{full}}^{\text{rct}}\right), \sigma^2 + (\sigma_\Gamma^+)^2 - 2\xi_\Gamma \right), \quad \text{as} \quad n_{\text{rct}} \to \infty \quad \text{and} \quad n_{\text{os}} \to \infty. \tag{7}$$

First, we observe that the mean of the LHS in Equation (6) converges to zero in probability:

$$\mathbb{E}_{\mathbb{P}^{\text{rct}}} \left[\frac{1}{\sqrt{n_{\text{rct}}}} \left| \sum_{X_i \in D_{\text{rct}}} \mu_\Gamma^+(X_i) - \hat{\mu}_\Gamma^+(X_i) \right| \right] \leq \sqrt{n_{\text{rct}}}\ \mathbb{E}_{\mathbb{P}^{\text{rct}}} \left[\left| \mu_\Gamma^+(X) - \hat{\mu}_\Gamma^+(X) \right| \right] \leq \sqrt{n_{\text{rct}}}\ \| \mu_\Gamma^+ - \hat{\mu}_\Gamma^+ \|_{L^2(\mathbb{P}^{\text{rct}})} = o_{\mathbb{P}^{\text{os}}}(1),$$

where in the last equality we have used $\| \mu_\Gamma^+ - \hat{\mu}_\Gamma^+ \|_{L^2(\mathbb{P}^{\text{rct}})} = O_{\mathbb{P}^{\text{os}}}(n_{\text{os}}^{-1/2})$ and $n_{\text{rct}} \ll n_{\text{os}}$. Next, we show that the variance term also converges to zero in probability, i.e.

$$\text{Var}_{\mathbb{P}^{\text{rct}}} \left[\frac{1}{\sqrt{n_{\text{rct}}}} \sum_{X_i \in D_{\text{rct}}} \mu_\Gamma^+(X_i) - \hat{\mu}_\Gamma^+(X_i) \right] \leq \mathbb{E}_{\mathbb{P}^{\text{rct}}} \left[\left(\mu_\Gamma^+(X) - \hat{\mu}_\Gamma^+(X) \right)^2 \right] = \| \mu_\Gamma^+ - \hat{\mu}_\Gamma^+ \|^2_{L^2(\mathbb{P}^{\text{rct}})} = o_{\mathbb{P}^{\text{os}}}(1),$$

and the statement in Equation (6) follows.

Therefore, by the consistency of $\hat{\sigma}^2, (\hat{\sigma}_\Gamma^+)^2$ and Slutsky's theorem, we have

$$\begin{aligned}
\lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\hat{T}_\Gamma^+ \leq z_{\alpha/2} \right) &= \lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\hat{\mu}_\Gamma^+ - \hat{\mu}}{\sqrt{(\hat{\sigma}_\Gamma^+)^2 + \hat{\sigma}^2 + 2\hat{\sigma}_\Gamma^+ \hat{\sigma}}} \leq z_{\alpha/2} \right) \\
&= \lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\sqrt{n_{\text{rct}}}(\hat{\mu}_\Gamma^+ - \hat{\mu})}{\sqrt{(\sigma_\Gamma^+)^2 + \sigma^2 + 2\sigma_\Gamma^+ \sigma}} \leq z_{\alpha/2} \right) \\
&\leq \lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\sqrt{n_{\text{rct}}}(\hat{\mu}_\Gamma^+ - \hat{\mu})}{\sqrt{(\sigma_\Gamma^+)^2 + \sigma^2 - 2\xi_\Gamma}} \leq z_{\alpha/2} \right),
\end{aligned}$$

where in the last line, we use Cauchy-Schwartz covariance inequality, i.e. $-\xi \leq |\xi_\Gamma| \leq \sigma_\Gamma^+ \sigma$. Finally, by asymptotic normality established in Equation (7), we conclude that

$$\lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\hat{T}_\Gamma^+ \leq z_{\alpha/2} \right) \leq \lim_{n_{\text{rct}}, n_{\text{os}} \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\sqrt{n_{\text{rct}}}(\hat{\mu}_\Gamma^+ - \hat{\mu}) - \mu_\Gamma^+ \left(\mathbb{P}_X^{\text{rct}}\right) + \mu \left(\mathbb{P}_X^{\text{rct}}, \mathbb{P}_{\text{full}}^{\text{rct}}\right)}{\sqrt{(\sigma_\Gamma^+)^2 + \sigma^2 - 2\xi_\Gamma}} \leq z_{\alpha/2} \right)$$
$$= \Phi(z_{\alpha/2}) = \alpha/2.$$

Proof of case $\diamond = \widetilde{\text{os}}$ Let $n = n_{\text{rct}} + n_{\text{os}}$ with fixed proportions, where $n_{\text{rct}} = \rho n$ and $n_{\text{os}} = (1-\rho)n$ for $\rho \in (0,1)$. Similarly to (1), by the central limit theorem and Lemma 3.2, it holds that

$$\sqrt{n} \sum_{(X_i, T_i, Y_i) \in D_{\text{rct}}} \left(\frac{Y_i T_i}{\pi} - \frac{Y_i(1-T_i)}{1-\pi} \right) w(X_i) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu\left(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}} \right), \sigma^2/\rho \right) \quad \text{as} \quad n \to \infty.$$

Then, from the asymptotic normality of $\hat{\mu}_\Gamma^+$ and the independence $\hat{\mu}_\Gamma^+ \perp\!\!\!\perp \hat{\mu}$, we have

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_\Gamma^+ \\ \hat{\mu} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} \mu_\Gamma^+ \\ \mu \end{pmatrix}, \begin{bmatrix} (\sigma_\Gamma^+)^2/(1-\rho) & 0 \\ 0 & \sigma^2/\rho \end{bmatrix} \right).$$

Hence, by the δ-technique with $h(X) = X_1 - X_2$, we get

$$\sqrt{n}\left(\hat{\mu}_\Gamma^+ - \hat{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu_\Gamma^+ \left(\mathbb{P}_X^{\widetilde{\text{os}}} \right) - \mu\left(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{rct}} \right), \frac{(\sigma_\Gamma^+)^2}{1-\rho} + \frac{\sigma^2}{\rho} \right) \quad \text{as} \quad n \to \infty.$$

Finally, from the consistency of $\hat{\sigma}^2, (\hat{\sigma}_\Gamma^+)^2$ and Slutsky's theorem, it holds that

$$\frac{\mu_\Gamma^+ - \hat{\mu}}{\sqrt{(\hat{\sigma}_\Gamma^+)^2 + \hat{\sigma}^2}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu_\Gamma^+ \left(\mathbb{P}_X^{\widetilde{\text{os}}} \right) - \mu\left(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}} \right), 1 \right) \quad \text{as} \quad n \to \infty.$$

As before, asymptotic normality implies that

$$\lim_{n \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\hat{T}_\Gamma^+ \leq z_{\alpha/2} \right) = \lim_{n \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\hat{\mu}_\Gamma^+ - \hat{\mu}}{\sqrt{(\hat{\sigma}_\Gamma^+)^2 + \hat{\sigma}^2}} \leq z_{\alpha/2} \right)$$
$$\leq \lim_{n \to \infty} \mathbb{P}_{H_0(\Gamma)} \left(\frac{\hat{\mu}_\Gamma^+ - \hat{\mu} - \mu_\Gamma^+ \left(\mathbb{P}_X^{\widetilde{\text{os}}} \right) + \mu \left(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}} \right)}{\sqrt{(\hat{\sigma}_\Gamma^+)^2 + \hat{\sigma}^2}} \leq z_{\alpha/2} \right)$$
$$= \Phi(z_{\alpha/2}) = \alpha/2.$$

A.2 Nested design

In a nested trial design, the randomized trial is embedded in a cohort of eligible people who are proposed to participate in the trial, but if they refuse, they are still included in the observational study. Two concrete examples of nested designs are the Women Health Initiative [1] and the recent study on Medicaid [14].

In contrast to Section 2, the nested design has an extra variable $S \in \{0,1\}$, which is a binary indicator for randomized trial participation. More formally, we observe i.i.d. samples from an underlying distribution $\mathbb{Q}_{\text{full}}$ over $(X, U, Y(0), Y(1), Y, S, T)$. Further, let $\mathbb{Q} := \mathcal{M}(\mathbb{Q}_{\text{full}})$ be the marginal distribution over (X, Y, S, T).

We can then write the marginal distributions over X of the (restricted) observational study and randomized trial as

$$\mathbb{P}_X^{\text{rct}}(X) = \mathbb{Q}\left(X \mid S = 1\right) \quad \text{and} \quad \mathbb{P}_X^{\widetilde{\text{os}}}(X) = \mathbb{Q}\left(X \mid S = 0, X \in \text{supp}(\mathbb{P}^{\text{rct}})\right).$$

This study design has a significant impact on the estimators previously introduced. In particular, the importance weights $w(X)$ can be estimated by pooling the observational study and the randomized trial as follows

$$\begin{aligned} w(X) &= \frac{\mathbb{P}^{\widetilde{\text{os}}}(X)}{\mathbb{P}^{\text{rct}}(X)} \\ &= \frac{\mathbb{Q}(X \mid S = 0, X \in \text{supp}(\mathbb{P}^{\text{rct}}))}{\mathbb{Q}(X \mid S = 1)} \\ &= \frac{\mathbb{Q}(S = 0 \mid X)}{\mathbb{Q}(S = 1 \mid X)} \frac{\mathbb{Q}(S = 1 \mid X \in \text{supp}(\mathbb{P}^{\text{rct}}))}{\mathbb{Q}(S = 0 \mid X \in \text{supp}(\mathbb{P}^{\text{rct}}))}, \end{aligned}$$

where the sampling probability $\mathbb{Q}(S \mid X)$ can be identified under a nested study design [11].

A.3 Limitations of MSM

We discuss here the intuition behind the tightness of $\hat{\Gamma}_{\texttt{LB}}$ in the infinite-sample limit, though it carries over to finite samples.

Without loss of generality, we focus on the lower bound derived from $\hat{\phi}_{\widetilde{\text{os}}}$ and assume that the unobserved confounding biases the average treatment effect upwards, i.e.

$$\mu^-_{\Gamma=1}(\mathbb{P}_X^{\widetilde{\text{os}}}) = \mu^+_{\Gamma=1}(\mathbb{P}_X^{\widetilde{\text{os}}}) \geq \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\widetilde{\text{os}}}),$$

where $\mu^-_{\Gamma=1}(\mathbb{P}_X^{\widetilde{\text{os}}})$ and $\mu^+_{\Gamma=1}(\mathbb{P}_X^{\widetilde{\text{os}}})$ are the IPW estimates of ATE on the (restricted) marginal distribution of the observational study. Further, we define the infinite-sample lower bound as

$$\Gamma_{\texttt{LB}} := \inf\left\{\Gamma : \lim_{n\to\infty} \hat{\phi}_{\widetilde{\text{os}}}(\Gamma, \alpha) = 0\right\},$$

which is, in words, the value of Γ such that the sensitivity bounds include the true ATE. Formally, it holds that

$$\begin{aligned} \Gamma_{\texttt{LB}} &= \inf\left\{\Gamma : \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}}) \in [\mu^+_\Gamma(\mathbb{P}_X^{\widetilde{\text{os}}}), \mu^-_\Gamma(\mathbb{P}_X^{\widetilde{\text{os}}})]\right\} \\ &= \inf\left\{\Gamma : \mu^+_\Gamma(\mathbb{P}_X^{\widetilde{\text{os}}}) = \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}}) \quad \text{or} \quad \mu^-_\Gamma(\mathbb{P}_X^{\widetilde{\text{os}}}) = \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}})\right\} \\ &= \inf\left\{\Gamma : \mu^-_\Gamma(\mathbb{P}_X^{\widetilde{\text{os}}}) = \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}})\right\}, \end{aligned}$$

where the second equality follows from the monotonicity of the sensitivity bounds and the last equality from the assumption of the bias direction. Since the sensitivity bounds are continuous and strictly increasing, the set is non-empty and contains one element.

The looseness of the lower bound can be characterized by $\Delta := \Gamma^\star - \Gamma_{\texttt{LB}}$: if $\Delta > 0$ even in the infinite-sample limit, the lower bound will not be tight. We discuss two interesting cases:

- If $\mathbb{P}_{\text{full}}^{\text{os}} = \underset{\tilde{\mathbb{P}}_{\text{full}} \in \mathcal{E}(\mathbb{P}^{\widetilde{\text{os}}}, \Gamma^\star)}{\text{argmax}} \mu(X, \tilde{\mathbb{P}}_{\text{full}})$, it holds that $\mu^+_{\Gamma^\star}(\mathbb{P}_X^{\widetilde{\text{os}}}) = \mu(\mathbb{P}_X^{\widetilde{\text{os}}}, \mathbb{P}_{\text{full}}^{\text{os}})$ and $\Delta = 0$. This case is achieved when $U = (Y(1), Y(0))$ (see Dorn and Guo [17] for a closed-form solution of the sensitivity bounds). Intuitively, the MSM does not place any assumptions on the form of the confounder, and the worst-case $\mathbb{P}_{\text{full}}^{\text{os}}$ is achieved when the confounder is equal to the potential outcomes.

- If $U \perp\!\!\!\perp (Y(1), Y(0))$, it holds that $\mu^{-}_{\Gamma=1}(\mathbb{P}^{\widetilde{\text{os}}}_X) = \mu^{+}_{\Gamma=1}(\mathbb{P}^{\widetilde{\text{os}}}_X) = \mu(\mathbb{P}^{\widetilde{\text{os}}}_X, \mathbb{P}^{\widetilde{\text{os}}}_{\text{full}})$. Hence, $\Delta = \Gamma^\star - 1$ and the lower bound can be arbitrarily loose.

B Experimental details

B.1 Synthetic experiments

We design the propensity scores such that the data distribution satisfies the MSM with true confounding strength $\Gamma^* = 5$. To do so, we define the *adversarial propensity score* as

$$e^+(X,U) = \begin{cases} \ell(X) & \text{if } U > t(X) \\ u(X) & \text{if } U \leq t(X) \end{cases}, \quad \text{where} \quad \ell(X) = \frac{e(X)}{e(X) + (1-e(X))\Gamma^\star}, \quad u(X) = \frac{e(X)}{e(X) + (1-e(X))/\Gamma^\star}$$

are respectively the lower and upper bounds on the full propensity score under the MSM. By choosing $t(X) = \frac{e(X)-\ell(X)}{u(X)-\ell(X)}$ in our data-generating process in Section 4.1 where $U \sim U(0,1)$, we ensure that $\mathbb{E}_{\mathbb{P}^{\text{os}}}[e^+(X,U) \mid X] = e(X)$. We note that this is different from [32, 35, 47] where they choose a fixed threshold $t(X) = 1/2$, resulting in a data distribution that does not satisfy the MSM. For all synthetic experiments, we estimate the propensity score using logistic regression. Further, we set $n_{\text{bootstrap}} = 100, \sigma_Y^2 = 0.1, \alpha = 0.05$, and report the mean and standard error over 20 runs.

For the test $\hat{\phi}_{\widetilde{\text{os}}}$, we use the sensitivity bound estimator QB [17], and fit the quantile function using quantile forest regression [41]. For the test $\hat{\phi}_{\text{rct}}$, we use the sensitivity bound estimator B-Learner [47]. We fit the quantile function using quantile forest regression [41], and the outcome model using a random forest regressor.

B.2 Semi-synthetic experiments

We provide the details of the semi-synthetic experiments in Section 4.1. Specifically, we describe the subsampling procedure used to generate a randomized trial and an observational study that satisfy our setting, along with additional information about the datasets employed.

B.2.1 Subsampling procedure

We now detail the procedure for constructing a randomized trial and multiple observational datasets for our semi-synthetic experiments. Given a large-scale real-world randomized trial D with covariates X_{all}, our objective is to create multiple observational datasets D_{os} that differ in the correlation $\rho_{u,y} = \frac{\text{Cov}_{\mathbb{P}^{\text{os}}_{\text{full}}}[Y(1),U]}{\sigma_{Y(1)}\sigma_U}$, i.e. $\mathbb{P}_{\text{inv}}$, but have the same confounding strength $\Gamma^\star$, i.e. $\mathbb{P}^{\text{os}}_{\text{cnf}}$. This setup allows us to separately understand the effect of $\rho_{u,y}$ on the power of the test. While we cannot directly intervene on $\mathbb{P}_{\text{inv}}(Y(1), Y(0)|X_{\text{all}})$ as it is intrinsic to the dataset, we can hide different $U \in X_{\text{all}}$ for each D_{os}, resulting in different $\mathbb{P}_{\text{inv}}(Y(1), Y(0)|U)$ and hence correlation coefficient $\rho_{u,y}$.

For each candidate hidden confounder U within X_{all}, we implement the following steps: First, we select a subset from D to construct our randomized trial dataset, D_{rct}, and remove U from X_{all}. Next, we *subsample* $D \setminus D_{\text{rct}}$ to generate a dataset D_{os} that belongs to $\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma^\star)$. We enforce this constraint by constructing a

Algorithm 2 Randomized Trial Rejection Sampling [37]

Inputs: $D \setminus D_{\text{rct}} = \{(X_i, U_i, Y_i, T_i)\}_{i=1}^n$; $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$, a function specified by the user; M a constant computed empirically.
Output: D_{os}.
$D_{\text{os}} \leftarrow D \setminus D_{\text{rct}}$
while true **do**
  for each unit i in D_{os} **do**
    Sample K_i uniform on $(0, 1)$
    if $K_i > \frac{\mathbb{P}^{\text{os}}_{\text{cnf}}(T=t_i|U_i)}{\hat{\mathbb{P}}^{\text{rct}}(T=t_i)M}$ **then**
      $D_{\text{os}} \leftarrow D_{\text{os}} \setminus \{(X_i, U_i, Y_i, T_i)\}$
    end if
  end for
  break if no units were discarded in the last iteration
end while
Remove U from D_{os}

specific propensity score $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$ (detailed in the sequel) and employing the subsampling procedure in Algorithm 2 using this $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$. Note that for simplicity, we choose a propensity score that does not depend on X; this is consistent with the graphical model in Figure 2.

We use different $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$ for continuous and binary confounders. We first define

$$\ell = \frac{\hat{\pi}}{\hat{\pi} + [1 - \hat{\pi}]\Gamma^\star}, \quad u = \frac{\hat{\pi}}{\hat{\pi} + [1 - \hat{\pi}]/\Gamma^\star}, \tag{8}$$

where we estimate $\hat{\pi} = \hat{\mathbb{P}}^{\text{rct}}(T = 1)$ from $D \setminus D_{\text{rct}}$. For a *continuous* confounder U positively correlated with $Y(1)$, we use the following full propensity score in Algorithm 2:

$$\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U) = \begin{cases} \ell & \text{if } U > Q_{\hat{q}^*}(U), \\ u & \text{if } U < Q_{\hat{q}^*}(U), \end{cases} \tag{9}$$

where $Q_q(U) = \inf\{z \in \mathbb{R} : \mathbb{P}^{\text{os}}_{\text{cnf}}(U \le z) \ge q\}$ is the q-th quantile of the marginal distribution of U. Since the propensity score does not depend on X, using $\hat{q}^\star = \frac{u-\hat{\pi}}{u-\ell}$ makes sure that

$$e(X) = \mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1|X) = \mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1) = \mathbb{E}_U\left[\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1|U)\mathbb{P}^{\text{os}}_{\text{cnf}}(U)\right] = \hat{\mathbb{P}}^{\text{rct}}(T = 1)$$

and hence the subsampled dataset, D_{os}, satisfies $\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma^\star)$. Note that this is equivalent to enforcing the same marginal propensity score before and after the subsampling. For a negatively correlated continuous confounder, we choose $\hat{q}^\star = \frac{\hat{\pi}-\ell}{u-\ell}$ and change the direction of the inequalities in Equation (9).

Further, for a positively correlated *binary* confounder, we use the following full propensity score in Algorithm 2:

$$\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U) = \begin{cases} \ell & \text{if } U = 1, \\ u & \text{if } U = 0. \end{cases} \tag{10}$$

By first subsampling $D \setminus D_{\text{rct}}$ such that

$$\hat{\mathbb{P}}^{\text{rct}}(U = 1) = \frac{u - \hat{\pi}}{u - \ell},$$

and then applying Algorithm 2 with the full propensity score from Equation (10), we again obtain $e(X) = \mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1) = \hat{\mathbb{P}}^{\text{rct}}(T = 1)$. For a negatively correlated binary confounder, we first enforce $\hat{\mathbb{P}}^{\text{rct}}(U = 1) = \frac{\hat{\pi}-\ell}{u-\ell}$ and swap ℓ for u in Equation (10), and vice versa.

The subsampling procedure in Algorithm 2 allows for the construction of observational datasets where the causal effect is identifiable non-parametrically, in contrast to previous approaches, such as [21], that do not guarantee identifiability after subsampling. In Algorithm 2, we note that $M = \frac{\max_{i\in\{1,\dots,n\}} \mathbb{P}^{\text{os}}_{\text{cnf}}(T=t_i|U_i)}{\min_{i\in\{1,\dots,n\}} \hat{\mathbb{P}}^{\text{rct}}(T=t_i)}$ in the limit satisfies $M \geq \sup_{T,U} \frac{\mathbb{P}^{\text{os}}_{\text{cnf}}(T|U)}{\mathbb{P}^{\text{rct}}(T)}$, which is required to ensure that the causal effect is identifiable in the subsampled dataset (see Theorem 3.2 in [37]). We empirically observe that the subsampling procedure approximately discards half of the instances from $D \setminus D_{\text{rct}}$.

B.2.2 Datasets details

We give additional details about the three datasets we use for the semi-synthetic experiments.

- **Hillstrom's MineThatData Email data** [27]. The Hillstrom dataset contains records of 64,000 customers who purchased within the last twelve months. They were part of an e-mail campaign to assess the effectiveness of distinct campaign strategies. Two treatment groups, "Men's" and "Women's" email campaigns, and a control group were established. Treatments were randomly assigned. Our analysis primarily focuses on a combined treatment group, which constitutes roughly 66% of the dataset. While the original dataset has different outcomes, we looked at the dollars spent in the two weeks post-campaign. The dataset provides data on recent purchase patterns (Recency), annual spending categories (History Segment) and values (History), merchandise type, either Mens (Mens) or Womens (Womens), geographical location via zip code (Zip Code), newcomer status (Newbie), and purchasing avenues (Channel). After subsampling, we end up with a randomized trial of size $n_{\text{rct}} = 7680$ and an observational dataset of size $n_{\text{os}} = 20500$. Assumption 2.3 is enforced by excluding urban zip codes in the trial.
- **VOTE dataset** [22]. The VOTE dataset studies the effect of social pressure on voting behaviors among Michigan's registered voters, focusing on those who voted in the prior election and met certain criteria. The primary outcome is a binary variable indicating whether the letter recipients voted. In this randomized trial, participants were allocated to a control group or one of four treatment groups. The treatment groups received distinct letters, each varying in social pressure intensity, aimed at encouraging voting. The most persuasive letter provided insight into the recipient's neighbors' voting patterns from the previous two elections and implied updates on neighbors' subsequent voting actions in future letters. Using the split in [55], we incorporated roughly 190,000 samples in the control group, and we kept the treatment group with the strongest letter, leaving about 38,000 samples. We retained preprocessed features like age, household size, gender, and two scores reflecting past voting habits and the voting patterns of neighbors. After subsampling, we end up with a randomized trial of size $n_{\text{rct}} = 10650$ and an observational dataset of size $n_{\text{os}} = 36800$. We discard households with more than 4 participants to enforce Assumption 2.3.
- **Tennessee STAR Project** [63]. The Tennessee STAR experiment, initiated in 1985, was a randomized study examining the impact of class size on students' standardized test scores, tracking them from kindergarten through third grade. Initially, students and teachers were randomly placed into class sizes, intending to maintain these conditions throughout the study. We follow the dataset preprocessing outlined in [34]. Their analysis concentrates on two conditions: small classes (13-17 students) and regular-sized classes (22-25 students). They used the class size at first grade as the treatment variable, observing 4,509 students. Their outcome aggregates scores from listening, reading, and math tests at the end of the first grade. After excluding students with missing values, the final sample consisted of 4,218 students: 1,805 in small classes (treatment) and 2,413 in regular-sized classes (control). The observed features for each student are gender, race, birth month, birthday, birth year, free lunch given or not and teacher ID. After subsampling, we end up with a randomized trial of size $n_{\text{rct}} = 600$ and

an observational dataset of size $n_{\text{os}} = 1800$. For the STAR project, we keep inner-city and suburban students but remove urban and rural ones to enforce Assumption 2.3.

We one-hot-encode all categorical features and standardize in $[0, 1]$ all continuous features.

Implementation We use QB [17] for the test $\hat{\phi}_{\widetilde{\text{os}}}$. For continuous outcomes (Hillstrom and STAR), we fit a random forest regressor for the quantile functions, while we leverage the closed-form solution for the quantiles in the binary case (VOTE). For the test $\hat{\phi}_{\text{rct}}$ we use the B-Learner [47], fitting respective random forest regressors for the quantiles, the outcome and the bounds models. For the binary case, we use the closed-form solution for the quantiles and fit the outcome and bounds models with XGBoost [4]. We always train a logistic regressor for the propensity score. We report mean and standard error over 15 runs and set $n_{\text{bootstrap}} = 200, \alpha = 0.05$ for all experiments.

B.3 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), originally 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 on trials testing three prevention strategies. Eligible women could choose to enrol 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 the prevention of 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 the prevention of 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.

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

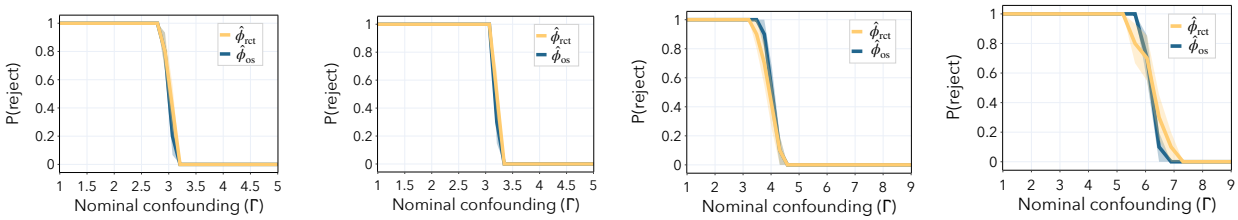
Data preprocessing We binarize a composite outcome, called the "global index", in our analysis, where Y = 1 if coronary heart disease or stroke was observed in the first seven years of follow-up, and Y = 0 otherwise. Note that Y = 0 could also occur from censoring. To establish treatment and control 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. 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 [49]. 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 train logistic regression for both outcome models and propensity score. We use as sensitivity bounds DVDS [18] for the test $\hat{\phi}_{\widetilde{\text{os}}}$, and B-Learner for the test $\hat{\phi}_{\text{rct}}$. We test for confounding in one direction, i.e. we only compute the test statistic $\hat{T}_{\Gamma}^{+}$. We set $n_{\text{bootstrap}} = 100$ and $\alpha = 0.05$ for all experiments.

C Additional Experiments

C.1 VOTE dataset

We present the experimental results with the VOTE dataset in Figure 4. Experiments were conducted with both weak and strong confounders, and under small and large sample regimes. We use the outcome Y as a strong confounder given the lack of a feature highly correlated with the outcome in the dataset. These results corroborate previous findings that higher correlated confounders and larger sample sizes lead to greater power of our test. In all scenarios, the performance of both tests closely aligns.

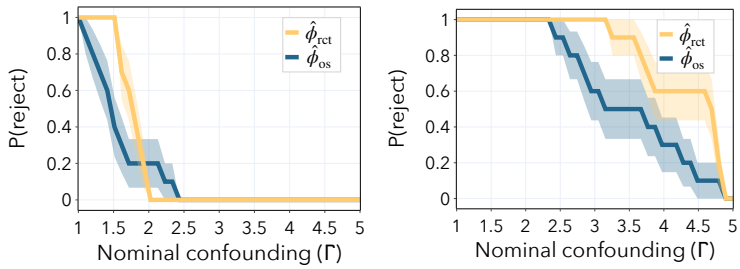


(a) Small sample "age" (b) Large sample "age" (c) Small sample Y (d) Large sample Y

Figure 4: Probability of rejection for different choices of Γ for the test for the VOTE dataset. For all the plots, the significance level is $\alpha = 0.05$ and $\Gamma^\star = 9$. (a)-(b) Weak confounder: "age". (a) small sample size: $n_{\text{rct}} = 3.2K, n_{\text{os}} = 11K$ and (b) large sample size: $n_{\text{rct}} = 10.6K, n_{\text{os}} = 36.8K$. (c)-(d) Strong confounder: outcome Y. (c) small sample size: $n_{\text{rct}} = 3.2K, n_{\text{os}} = 11K$ and (d) large sample size: $n_{\text{rct}} = 10.6K, n_{\text{os}} = 36.8K$.

C.2 Tennessee STAR Project

We present the experimental results with the STAR Project in Figure 5. Experiments were conducted with both weak and strong confounders with the full dataset. We do not run experiments with a small sample size since the STAR dataset already represents a small sample regime. These results corroborate previous findings that higher correlated confounders lead to greater power.



(a) Weak confounder: "free lunch" (b) Strong confounder: Y

Figure 5: Probability of rejection for different choices of Γ for the test for the STAR Project. For all the plots, the significance level is $\alpha = 0.05$ and $\Gamma^\star = 5$. We use the original sample sizes $n_{\text{rct}} = 600, n_{\text{os}} = 1.8K$. (a) weak confounder: "free lunch" (b) strong confounder: outcome Y.