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

Piersilvio De Bartolomeis* **Javier Abad*** **Konstantin Donhauser** **Fanny Yang**
ETH Zurich and ETH AI Center

Abstract

In the era of fast-paced precision medicine, observational studies play a major role in properly evaluating new drugs in clinical practice. Yet, unobserved confounding can significantly compromise causal conclusions from observational data. We propose a novel strategy to quantify unobserved confounding by leveraging randomized trials. First, we design a statistical test to detect unobserved confounding with strength above a given threshold. 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 setting.

1 Introduction

Monitoring the performance of a new drug after its approval is crucial, a process commonly referred to as *postmarketing surveillance* (Vlahović-Palčevski and Mentzer, 2011). This task is especially important for rare diseases and targeted therapies, where treatment approval often relies on a small randomized trial that is not representative of the target population (Franklin et al., 2019; FDA, 2023). In such cases, there is significant uncertainty about the treatment effectiveness, and thus, it is imperative to gather more evidence using observational data from clinical practice (Platt et al., 2018; Klonoff, 2020).

Preliminary work. Under review by AISTATS 2024. Do not distribute.

However, unobserved confounding usually prevents reliable causal conclusions from observational data. Although this problem cannot be addressed without further assumptions, sensitivity analysis offers a partial solution (Cornfield et al., 1959). This field studies how a specific strength of unobserved confounding might alter conclusions but does not address which strengths are plausible nor whether unobserved confounding is present. As a result, epidemiologists often rely on heuristic judgments to evaluate the soundness of an observational study.

In postmarketing surveillance, we often have access to a randomized trial free of confounding by design. This additional data enables the use of different strategies for tackling the problem of unobserved confounding. One common approach is to combine the estimators from randomized trials and observational studies (see Colnet et al. (2020); Brantner et al. (2023) for a survey of methods). However, this approach crucially relies on some prior knowledge of the confounding bias structure and can lose effectiveness when there is substantial confounding bias in the observational study (Chen et al., 2021; Oberst et al., 2022).

Our work proposes an alternative paradigm to leverage randomized trials, that is, to test and quantify the unobserved confounding strength. In particular, if strong confounding is detected, epidemiologists can take proactive measures to correct it. Most directly, they can identify and incorporate important covariates into the study design if they were initially overlooked (Dreyer, 2018). On the other hand, if small confounding is detected, epidemiologists can continue their analysis (see Figure 1 for an illustration of the pipeline). Our approach is closely related to a line of work that proposes tests for the *presence* of hidden confounding (Viele et al., 2014; Hussain et al., 2022, 2023; Morucci et al., 2023). However, these tests cannot quantify the strength of unobserved confounding and may reject even if the strength is minimal.

In this paper, we use the marginal sensitivity model to quantify the unobserved confounding strength (Tan,

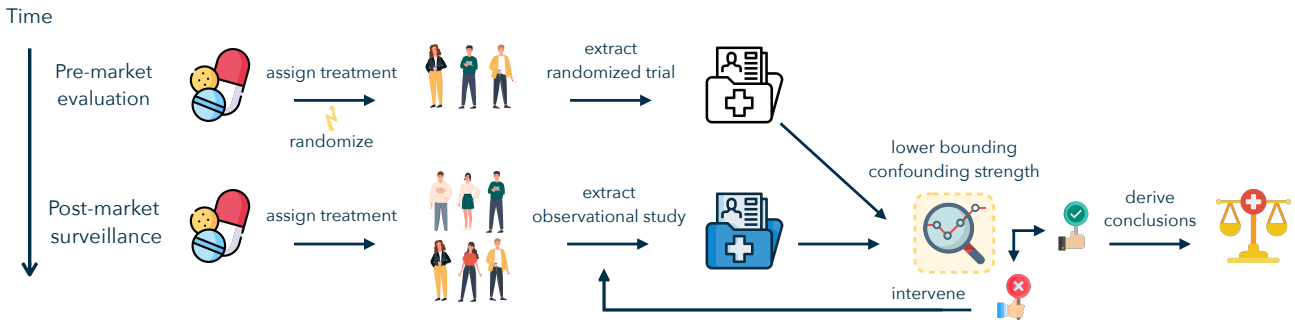


Figure 1: An illustrative example of the drug regulatory process: our lower bound allows taking proactive measures to address the unobserved confounding issue.

2006). In Section 4, we introduce the first test to detect unobserved confounding with strength above a certain threshold. Crucially, our test provides a lower bound on the unobserved confounding strength. In Section 5, we evaluate the finite-sample validity and power of our test on several synthetic and semi-synthetic datasets. In Section 6, we demonstrate in a real-world example how our lower bound can be used to derive conclusions that align with epidemiological understanding. Finally, in Section 7, we discuss the limitations of our method and identify directions for future work.

2 Related work

Several works have introduced tests to detect unobserved confounding. Viele et al. (2014); Morucci et al. (2023) propose similar tests that check if there is a statistically significant difference between average treatment effect estimates obtained from randomized trials and observational studies. Hussain et al. (2022) explore differences in group-level treatment effects, while Hussain et al. (2023) focus on differences in the conditional average treatment effects.

Similarly, De Luna and Johansson (2014); Donald et al. (2014) propose an average treatment effect test but exploit access to instrumental variables rather than randomized trials. Karlsson and Krijthe (2022) leverage access to multiple environments and test conditional independences to detect confounding. Further, Lipsitch et al. (2010); Sofer et al. (2016) exploit negative control outcomes for detecting unobserved confounding.

In contrast to our test, these works have a significant limitation: they cannot quantify the 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 is too restrictive.

3 Setting and notation

We consider the Neyman-Rubin potential outcomes model with a binary treatment. More formally, let $\mathbb{P}_{\text{full}}$ denote a distribution over $(X, U, Y(0), Y(1), Y, T)$, where $(X, U) \in \mathbb{R}^d \times \mathbb{R}^k$ is a vector of confounders, $(Y(0), Y(1))$ are real-valued bounded potential outcomes and $T \in \{0, 1\}$ is a binary treatment indicator.

Throughout the paper, we consider two distributions, namely $\mathbb{P}^{\text{rct}}_{\text{full}}$ and $\mathbb{P}^{\text{os}}_{\text{full}}$, which represent a randomized trial and an observational study. Observe that we can factorize the full distributions for both $\diamond \in \{\text{rct}, \text{os}\}$ as

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

where $\mathbb{P}_{\text{det}}$ is deterministically given by $Y = Y(1)T + Y(0)(1-T)$, $\mathbb{P}_{\text{inv}}$ is invariant to the study design, and $\mathbb{P}^{\diamond}_{\text{cnf}}$ depends on it. In Figure 2, we depict the corresponding graphical model. In particular, note that this model can capture the essence of the potential outcomes framework while being more general, e.g. it can allow for distribution shift, and one can easily incorporate many standard assumptions in the literature (Section 3.1). We further use the shorthand $\mathcal{M}(\mathbb{P}^{\diamond}_{\text{full}})$ to denote the marginal distribution of X, Y, T under $\mathbb{P}^{\diamond}_{\text{full}}$, and write

$$\mathbb{P}^{\diamond} := \mathcal{M}(\mathbb{P}^{\diamond}_{\text{full}}) = \mathbb{P}^{\diamond}_{X,Y,T}, \quad \text{for } \diamond \in \{\text{rct}, \text{os}\}.$$

During inference time, we have access to a randomized trial $D_{\text{rct}} = \{(X_i, Y_i, T_i)\}_{i=1}^{n_{\text{rct}}}$, sampled from the distribution $\mathbb{P}^{\text{rct}}$ and an observational study $D_{\text{obs}} = \{(X_i, Y_i, T_i)\}_{i=1}^{n_{\text{os}}}$, sampled from $\mathbb{P}^{\text{os}}$.

3.1 Further details and assumptions

First, in order to leverage the randomized trial to detect confounding in the observational study, some parts of the distribution need to be shared across the two datasets. The minimum assumption we require for comparing the two studies is the following property of the conditional average treatment effect (CATE).

Assumption 3.1 (Transportability)**.** *The conditional average treatment effect remains invariant across studies, that is*

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

In the literature, this property is also called *transportability of CATE* (Colnet et al., 2020) and is a weaker assumption than sample ignorability for treatment effects (Kern et al., 2016) or conditional ignorability (Hartman et al., 2021).

As opposed to the invariant conditional distribution $\mathbb{P}_{\text{inv}}$, the confounding structure reflected in $\mathbb{P}_{\text{cnf}}$ differs in the different datasets: in the observational study, T can depend on X, U arbitrarily i.e. $\mathbb{P}^{\text{os}}_{\text{cnf}} = \mathbb{P}^{\text{os}}_{T|X,U}\,\mathbb{P}^{\text{os}}_{X,U}$, while for the randomized trial we assume internal validity.

Assumption 3.2 (Internal validity)**.** *The treatment is assigned independent of the covariates and the potential outcomes, that is, in our setting,*

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

Further, our factorization (1) allows for distribution shifts in the marginal distributions over the observed and hidden confounders X, U. For our test, we require an additional assumption

Assumption 3.3 (Support inclusion)**.** *The support of the randomized trial is included in the support of the observational study, i.e.*

$$\text{supp}(\mathbb{P}^{\text{rct}}_{X}) \subset \text{supp}(\mathbb{P}^{\text{os}}_{X}).$$

This assumption is strictly weaker than the positivity of trial participation, see e.g. Andrews and Oster (2017); Hartman et al. (2015); Nie et al. (2021); Stuart et al. (2011); Colnet et al. (2022) and aligns with how observational studies are constructed in postmarketing surveillance (Franklin et al., 2019; Schurman, 2019).

3.2 Sensitivity analysis

Sensitivity analysis is the most common way to address unobserved confounding in observational data.

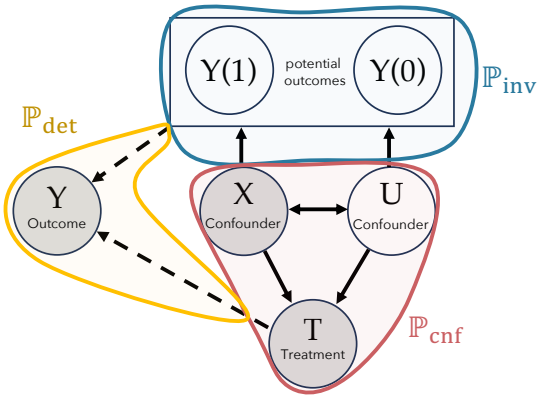


Figure 2: Graphical model for the Neyman-Rubin potential outcome framework. $\mathbb{P}_{\text{cnf}}$ is the part of the distribution which changes across studies. Observed variables are colored in shades of grey.

In particular, this approach estimates an interval for the treatment effect that depends on an assumed *confounding strength* Γ of $\mathbb{P}_{\text{cnf}}$. Throughout this paper, we define the confounding strength using the widely accepted marginal sensitivity model (MSM), first introduced in Tan (2006).

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

$$\Gamma^{-1} \leq \frac{\mathbb{P}_{\text{cnf}}(T = 1 \mid X, U)}{\mathbb{P}_{\text{cnf}}(T = 0 \mid X, U)} \Big/ \frac{\mathbb{P}_{\text{cnf}}(T = 1 \mid X)}{\mathbb{P}_{\text{cnf}}(T = 0 \mid X)} \leq \Gamma.$$

As a consequence, if we assume a confounding strength of Γ on $\mathbb{P}_{\text{cnf}}$, we can define a set of joint distributions $\widetilde{\mathbb{P}}_{\text{full}}$ that are compatible with $\mathbb{P}^{\text{os}}$.

Definition 3.1 (Marginal sensitivity set)**.** *Given an observational 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 $\widetilde{\mathbb{P}}_{\text{full}}$, as*

$$\mathcal{E}(\mathbb{P}^{\text{os}}, \Gamma) := \{\widetilde{\mathbb{P}}_{\text{full}} = \mathbb{P}_{\text{det}}\tilde{\mathbb{P}}_{\text{inv}}\tilde{\mathbb{P}}_{\text{cnf}} : \tilde{\mathbb{P}}_{\text{cnf}} \in \mathcal{E}(\Gamma) \text{ and } \mathcal{M}(\mathbb{P}_{\text{det}}\tilde{\mathbb{P}}_{\text{inv}}\tilde{\mathbb{P}}_{\text{cnf}}) = \mathbb{P}^{\text{os}}\}. \quad (2)$$

In other words, this set contains all full joint distributions that could result in the observational distribution $\mathbb{P}^{\text{os}}$. Under the MSM, it is possible to partially identify the (conditional) treatment effect. More concretely, we define the following sensitivity analysis bounds.

Definition 3.2 (Sensitivity bounds)**.** *We define the conditional average treatment effect as*

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

and the upper and lower bounds on CATE under the

MSM as

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

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 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 number, depending on its argument. Several estimators for these population quantities have recently emerged in the literature for the CATE bounds (Kallus et al., 2019; Jesson et al., 2021; Oprescu et al., 2023) and for the ATE bounds (Zhao et al., 2019; Dorn et al., 2021; Dorn and Guo, 2022).

4 Estimating confounding strength

We would like to know 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 $\mathbb{P}^{\text{os}}$. *Binary testing* for unobserved confounding has been recently studied (see Section 2). However, in practice, we would like to accept $H_0(\Gamma)$ if the unobserved confounding does not compromise causal conclusions and, hence, these previous approaches are overly stringent.

In this section, we propose the first test, to the best of our knowledge, for a nominal confounding strength 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 nominal 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 4.1. *Under Assumption 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 $\mathbb{P}_X$, 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 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. □

In the following sections, we propose estimates for the average treatment effect under two different target populations. Further, we design an asymptotically valid test at significance level α, which can then be directly used to establish an asymptotically valid lower bound on the unobserved confounding strength.

4.1 Statistical tests for $H_0(\Gamma)$

Estimating the ATE We now discuss how $\mu(\mathbb{P}^\diamond_X, \mathbb{P}^\diamond_{\text{full}})$ can be estimated using data from the randomized trial, where internal validity holds. We first need to define a target population $\mathbb{P}^\diamond_X$ on which to estimate the ATE. In particular, the following lemma shows how the choice of $\mathbb{P}^{\text{rct}}_X$ and $\mathbb{P}^{\widetilde{\text{os}}}_X$ with $\mathbb{P}^{\widetilde{\text{os}}} := \mathbb{P}^{\text{os}} \mid X \in \text{supp}(\mathbb{P}^{\text{rct}})$ has favorable properties.

Lemma 4.2. *For $\diamond \in \{\text{rct}, \widetilde{\text{os}}\}$, under Assumptions 3.1, 3.2 and 3.3, we have*

$$\mu(\mathbb{P}^\diamond_X, \mathbb{P}^\diamond_{\text{full}}) = \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[Y\left(\frac{T}{\pi} - \frac{(1-T)}{1-\pi}\right) w(X)\right],$$

where $w(X) := \frac{\mathbb{P}^\diamond(X)}{\mathbb{P}^{\text{rct}}(X)}$.

Lemma 4.2 is a well-known result in the transportability literature, see Cole and Stuart (2010); Colnet et al. (2023). Essentially, it establishes that, when the distribution shift between $\mathbb{P}^{\text{rct}}_X$ and $\mathbb{P}^\diamond_X$ can be corrected, we can identify and estimate the ATE under $\mathbb{P}^\diamond_X$.

Estimating the sensitivity interval We discuss how $\mu^-_\Gamma(\mathbb{P}^\diamond_X), \mu^+_\Gamma(\mathbb{P}^\diamond_X)$ can be estimated using data from both the observational study and the target population $\mathbb{P}^\diamond_X$. Here, the approach varies based on the target population. For $\mathbb{P}^\diamond = \mathbb{P}^{\text{rct}}$, we estimate CATE sensitivity bounds from observational data and average them over the target population. Specifically, we use the B-Learner (Oprescu et al., 2023). For $\mathbb{P}^\diamond = \mathbb{P}^{\widetilde{\text{os}}}$, we have two options: either estimate the CATE sensitivity bounds and average them, or directly estimate the ATE sensitivity bounds over the target population. In our experiments, we estimate ATE sensitivity bounds using either the DVDS (Dorn et al., 2021) or the QB estimator (Dorn and Guo, 2022). These methods yield estimates that are valid, sharp, and efficient under more general conditions than other existing methods. Nevertheless, our testing procedure is

Algorithm 1 Test for detecting unobserved confounding under $\mathbb{P}^\diamond$

1: **Input:** $\diamond \in \{\text{rct}, \widetilde{\text{os}}\}$, $D_{\text{rct}}, D_{\text{obs}}$, 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 transport the bounds on the target population $\mathbb{P}^\diamond$:

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

4: Compute the test statistics:

$$\hat{T}^+_\Gamma = \frac{\hat{\mu}^+_\Gamma - \hat{\mu}}{\sqrt{(\hat{\sigma}^+_\Gamma)^2 + \hat{\sigma}^2 + 2\hat{\sigma}^+_\Gamma \hat{\sigma}}}, \quad \hat{T}^-_\Gamma = \frac{\hat{\mu} - \hat{\mu}^-_\Gamma}{\sqrt{(\hat{\sigma}^-_\Gamma)^2 + \hat{\sigma}^2 + 2\hat{\sigma}^-_\Gamma \hat{\sigma}}}.$$

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.

agnostic with respect to the sensitivity bound estimator, allowing for various options to be adopted.

Two statistical tests In Algorithm 1, we outline our testing procedure which can be instantiated for $\diamond \in \{\text{rct}, \widetilde{\text{os}}\}$. 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 4.1 (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 Assumption 3.1, 3.2 and 3.3, we have, for $H_0(\Gamma)$,*

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

$$\|\mu^-_\Gamma - \hat{\mu}^-_\Gamma\|_2 = o_P(n_{\text{os}}^{-1/2}), \quad \|\mu^+_\Gamma - \hat{\mu}^+_\Gamma\|_2 = o_P(n_{\text{os}}^{-1/2}),$$

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

(2) If $\hat{\mu}^+_\Gamma$ and $\hat{\mu}^-_\Gamma$ are consistent estimators of the ATE sensitivity 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 refer the reader to Appendix A.1.2 for a complete proof. In essence, we propose two tests that work under different assumptions: $\hat{\phi}_{\text{rct}}$ relies on a consistent estimate of the CATE sensitivity bounds, while $\hat{\phi}_{\widetilde{\text{os}}}$ requires an estimate of the importance weights $w(X)$.

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 or the classes are imbalanced), but the weights $w(X)$ can be identified and vice versa. The importance weights can be estimated for example when the observational study and the randomized trial adhere to a nested trial design (e.g. in Olschewski and Scheurlen (1985); Olschewski et al. (1992); Choudhry (2017)). See Appendix A.2 for a discussion on how the importance weights are estimated in this setting. In addition, $\hat{\phi}_{\widetilde{\text{os}}}$ benefits from large observational studies as the variance of the estimates $\hat{\mu}^+_\Gamma$ and $\hat{\mu}^+_\Gamma$ depends only on the sample size of the observational study.

4.2 A lower bound on confounding strength

In certain contexts, it is not sufficient for epidemiologists to detect the existence of unobserved confounding. They also require a measure of its strength, especially when collecting data is expensive and discarding a study because of minimal confounding is a problem.

We show how our test can provide 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, our aim is to find a quantity that with high probability is a lower bound for the true confounding strength $\Gamma^\star$. We first fix a test $\hat{\phi} \in \{\hat{\phi}_{\text{rct}}, \hat{\phi}_{\widetilde{\text{os}}}\}$, and recall that $\hat{\phi}(\Gamma, \alpha)$ is a deterministic continuous function[1]. Hence, to obtain a lower bound for a fixed significance level

[1]We fix the randomness in the bootstrap estimate of the variance for all Γ.

α, we can compute

$$\hat{\Gamma}_{\text{LB}} = \inf_{\Gamma}\{\Gamma : \hat{\phi}(\Gamma, \alpha) = 0\} \tag{3}$$

which is, in words, the smallest Γ such that the test accepts. We show in the following proposition that $\hat{\Gamma}_{\text{LB}}$ is then indeed a valid lower bound for $\Gamma^\star$.

Proposition 4.2 (Multiple testing)**.** *Let $\hat{\Gamma}_{LB}$ be as in Equation (3) for a fixed significance level α. Then, under the setting described in Section 3, $\hat{\Gamma}_{LB}$ is an asymptotically valid lower bound, i.e.*

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

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

$$\begin{aligned}\mathbb{P}(\hat{\Gamma}_{\text{LB}} \geq \Gamma^\star) &= \mathbb{P}(\cup_{\Gamma \leq \Gamma^\star}\{\hat{\phi}(\Gamma, \alpha) = 1\}) \\ &\leq \mathbb{P}(\hat{\phi}(\Gamma^\star, \alpha) = 1) \leq \alpha + o(1),\end{aligned}$$

where the last inequality follows from the asymptotic validity of the test in Proposition 4.1. □

5 Synthetic and semi-synthetic experiments

In this section, we evaluate the test and the derived lower bound $\hat{\Gamma}_{\text{LB}}$ in finite-sample synthetic and semi-synthetic experiments. In particular, we study how these change with different $\mathbb{P}_{\text{inv}}$ for fixed $\Gamma^\star$ (i.e. $\mathbb{P}_{\text{cnf}}$) and with the size of the observational study.

We postulate that, for a fixed $\Gamma^\star$, the tightness of the sensitivity bounds (and hence the power of the test) increases with the correlation between the unobserved confounder and the potential outcomes measured by the correlation coefficient

$$\rho_{u,y} = \frac{\text{Cov}_{\mathbb{P}^{\text{pos}}_{\text{full}}}[Y(1), U]}{\sigma_Y \sigma_U}. \tag{4}$$

In the case of adversarial confounding, where $U = Y$, the correlation coefficient $\rho_{u,y} = 1$ and $\hat{\Gamma}_{\text{LB}} = \Gamma^\star$. Conversely, when U is independent of Y, $\rho_{u,y} = 0$ and $\hat{\Gamma}_{\text{LB}} = 1$. Hence, we expect the power of the test to increase as the correlation coefficient approaches 1. For additional details, we refer the reader to Appendix A.3.

Moreover, studying the test's behavior as the observational study grows is of practical relevance. In realistic scenarios, the number of samples in a randomized trial is unlikely to increase; however, an observational study has the potential to grow over time.

5.1 Datasets

Synthetic distribution We first benchmark our test with a synthetic distribution similar to the one used by Yadlowsky et al. (2018); Jin et al. (2023), where treatment selection and outcome are controlled. Here, both propensity score (and corresponding $\Gamma^\star$) and 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 $U \sim \text{Unif}(0, 1)$, and the observed covariates

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

Further, for the observational distribution, we assume that the conditional distribution of the treatment T given X, U is a Bernoulli and satisfies the marginal sensitivity model with $\Gamma^\star$. Specifically, we fix the marginal propensity score as

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

and design the true propensity score $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid X, U)$ such that it is consistent with $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid X)$. For the randomized control trial we choose

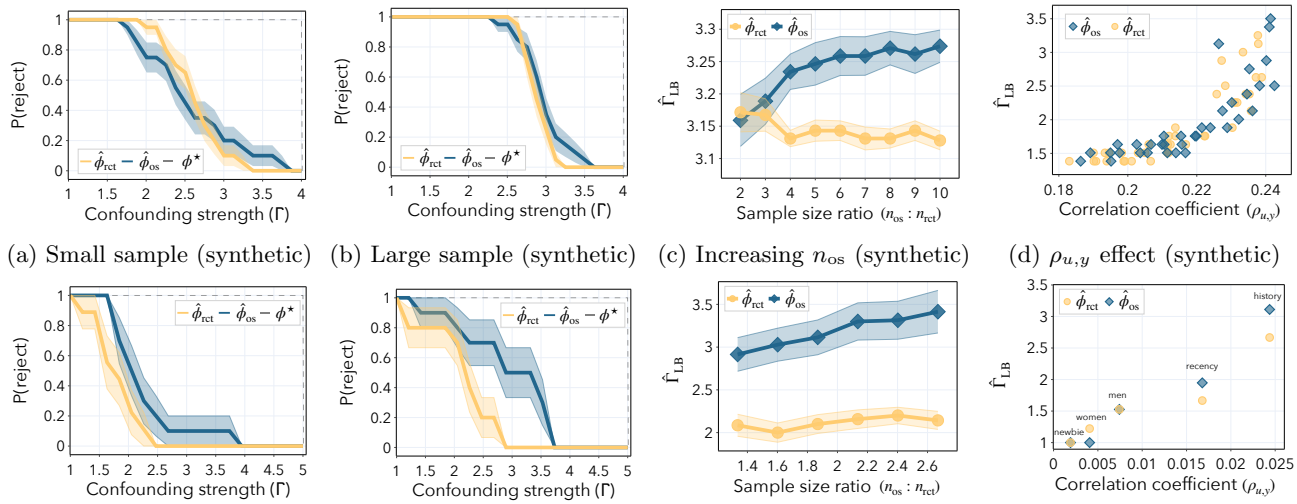
$$\pi = \mathbb{P}^{\text{rct}}_{\text{cnf}}(T = 1 | X, U) = 1/2.$$

We refer the reader to Appendix B.1 for complete experimental details.

Semi-synthetic datasets We further assess our test using three real-world randomized trials: Hillstrom's MineThatData Email data (Hillstrom, 2008), the Tennessee STAR study (Word et al., 1990) and the VOTE dataset (Gerber et al., 2008). Due to space constraints, we present experiments solely for Hillstrom's data and refer to Appendix C.2 for the other two datasets. Hillstrom's data aimed to identify the effect of an email campaign on the dollars spent by the recipients in the following two weeks.

We first select a small subset of the original trial as our randomized trial D_{rct} and refer to the remaining subset as $\hat{D}_{\text{obs}}$. We can then subsample multiple observational studies from $\hat{D}_{\text{obs}}$ sharing a fixed true strength $\Gamma^\star$ i.e. $\mathbb{P}_{\text{cnf}}$, but with a varying correlation between the hidden confounder U and outcome Y, i.e. $\mathbb{P}_{\text{inv}}$.

Let us denote the vector of all observed covariates as X_{all}. While we cannot directly intervene on



(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 $\alpha = 0.05$ and $\phi^\star$ denotes the oracle test which rejects for $\Gamma < \Gamma^\star$. First row with *synthetic experiment* with $\Gamma^\star = 4$: Probability of rejection for different choices of Γ for the test for (a) small sample size: $n_{\text{rct}} = 2K, n_{\text{os}} = 2K$ and (b) large sample size: $n_{\text{rct}} = 5K, n_{\text{os}} = 2K$. $\hat{\Gamma}_{\text{LB}}$ for (c) increasing sample size of the observational study with $n_{\text{rct}} = 10K$ and (d) for increasing correlation coefficient; $n_{\text{rct}} = 10K, n_{\text{os}} = 50K$. Second row with *semi-synthetic* Hillstrom dataset choosing $\Gamma^\star = 5$ and using "history" as unobserved confounder (except in (h)): Probability of rejection for different Γ : 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 of the unobserved confounder.

$\mathbb{P}_{\text{inv}}(Y(1), Y(0)|X_{\text{all}})$ as it is intrinsic to the dataset, we can generate multiple observational studies by partitioning X_{all} into a hidden confounder U and observed X in different ways. For a given partitioning $X_{\text{all}} = (U, X)$, the resulting D_{obs} will have a certain $\mathbb{P}_{\text{inv}}(Y(1), Y(0)|U)$ and hence correlation coefficient $\rho_{u,y}$. For this choice of U, we then enforce a true propensity score $\mathbb{P}^{\text{os}}_{\text{cnf}}(T = 1 \mid U)$ that satisfies $\mathcal{E}(\Gamma^\star)$ by subsampling $\hat{D}_{\text{obs}}$. Finally, we remove U to construct D_{obs}. Our approach is a variation of the methods presented by Keith et al. (2023); Gentzel et al. (2021) (see further details in Appendix B.2).

In order for both tests to be computable, we enforce Assumption 3.3 where we reduce support of the randomized trial implicitly by excluding urban zip codes.

5.2 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 semi-synthetic.

Effect of observational study sample size First, we observe in Figure 3a-3b and Figure 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 Figure 3c-3g, the lower bound $\hat{\Gamma}_{\text{LB}}$ varies with the sample size of the observational study. We confirm that the $\hat{\phi}_{\widetilde{\text{os}}}$ derives greater benefits from a larger sample size than $\hat{\phi}_{\text{rct}}$, as discussed in Section 4.1.

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 $\hat{\Gamma}_{\text{LB}}$. We study the effect of increasing the outcome-confounder correlation (Equation 4). Specifically, we generate observational datasets with a constant $\Gamma^\star$ but varying $\rho_{u,y}$, and report $\hat{\Gamma}_{\text{LB}}$. For the synthetic experiments in Figure 3d, we plot $\hat{\Gamma}_{\text{LB}}$ for both tests against $n = 40$ distinct values of $\sigma_2^Y \sim \text{Unif}[0, 1]$. For the semi-synthetic experiments in Figure 3h, we depict $\hat{\Gamma}_{\text{LB}}$ for both tests across different hidden features U. Both plots confirm our hypothesis that higher $\rho_{u,y}$ correlates with higher power in terms of $\hat{\Gamma}_{\text{LB}}$.

6 Real-world experiments

Linking back to the pipeline in Figure 1, we demonstrate how epidemiologists can use the lower bound

$\hat{\Gamma}_{\text{LB}}$ to successfully differentiate between studies with significant confounding to those with small confounding. Specifically, we propose comparing $\hat{\Gamma}_{\text{LB}}$ with a critical value of Γ, also derived from the available data

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

In essence, $\hat{\Gamma}_{\text{CT}}$ represents the strength of confounding for which sensitivity analysis includes both positive and negative values, thereby invalidating the results of the study. We flag an observational study as confounded if $\hat{\Gamma}_{\text{LB}}$ exceeds the critical value, i.e.

$$\psi_{\text{sens}} := \mathbb{I}\{\hat{\Gamma}_{\text{LB}} > \hat{\Gamma}_{\text{CT}}\}. \tag{5}$$

We compare our decision with a procedure based on a binary test that is valid by design

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

In contrast to our procedure, the output of the binary test flags an observational study if any strength of confounding is detected. Note that choosing any more sophisticated binary test in the literature with more power would only *exacerbate* this gap.

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 (Vandenbroucke, 2009) based on the findings of the Women's Health Initiative (WHI) study (Anderson et al., 2003). The WHI study included a (randomized) Postmenopausal Hormone Therapy trial and an observational study that examined the impact of HRT on various cancers, cardiovascular events, and fractures. 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. We now present evidence that our procedure can yield the same epidemiological conclusions and avoids issuing false alarms.

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 (5) to both the original dataset, which includes all patients (i.e. $t = 20$), 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 $\hat{\Gamma}_{\text{LB}}$ 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 confounding ($t = 0$) and with large confounding ($t = 20$). For coronary heart disease, both algorithms flag the study as confounded when strong unobserved confounding is present ($t = 20$). However, when minimal unobserved confounding is present ($t = 0$), our test does not flag the study, while ψ_{bin} does. This difference underscores our test's capability to distinguish between small and large unobserved confounding, thereby addressing a limitation in the flagging procedures based on existing testing methods.

	Stroke		Coronary heart disease	
	$t = 0$	$t = 20$	$t = 0$	$t = 20$
$\hat{\Gamma}_{\text{CT}}$	1.017	1.172	1.017	1.164
$\hat{\Gamma}_{\text{LB}}$	1.052	1.207	1.009	1.224
ψ_{bin}	1	1	1	1
ψ_{sens}	1	1	0	1

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

In the case of stroke, both ψ_{sens} and ψ_{bin} correctly flag the observational study for both confounding strengths, even when we adjust for the time since the start of treatment. This finding aligns with experts (Prentice et al., 2005) suggesting that additional hidden confounding factors for stroke are still present.

7 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 on subgroups that cancel 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.