Falsification before Extrapolation in Causal Effect Estimation

Zeshan Hussain*
MIT CSAIL & IMES
Cambridge, MA
zeshanmh@mit.edu

Michael Oberst*
MIT CSAIL & IMES
Cambridge, MA
moberst@mit.edu

Ming-Chieh Shih*
National Dong Hwa University
Hualien, Taiwan
mcshih@gms.ndhu.edu.tw

David Sontag
MIT CSAIL & IMES
Cambridge, MA
dsontag@csail.mit.edu

Abstract

Randomized Controlled Trials (RCTs) represent a gold standard when developing policy guidelines. However, RCTs are often narrow, and lack data on broader populations of interest. Causal effects in these populations are often estimated using observational datasets, which may suffer from unobserved confounding and selection bias. Given a set of observational estimates (e.g. from multiple studies), we propose a meta-algorithm that attempts to reject observational estimates that are biased. We do so using *validation effects*, causal effects that can be inferred from both RCT and observational data. After rejecting estimators that do not pass this test, we generate conservative confidence intervals on the *extrapolated* causal effects for subgroups not observed in the RCT. Under the assumption that at least one observational estimator is asymptotically normal and consistent for both the validation and extrapolated effects, we provide guarantees on the coverage probability of the intervals output by our algorithm. To facilitate hypothesis testing in settings where causal effect transportation across datasets is necessary, we give conditions under which a doubly-robust estimator of group average treatment effects is asymptotically normal, even when flexible machine learning methods are used for estimation of nuisance parameters. We illustrate the properties of our approach on semi-synthetic and real world datasets, and show that it compares favorably to standard meta-analysis techniques.

1 Introduction

Policy guidelines often rely on conclusions from Randomized Controlled Trials (RCTs), whether considering treatment decisions in healthcare, classroom interventions in education, or social programs in economics [31, 13, 43]. In healthcare, when a target population has reasonable overlap with the inclusion criteria of RCTs, current clinical treatment guidelines rely primarily on RCTs [23, 22]. For target populations not well-represented in RCTs, observational studies are often used to infer treatment effects. However, different observational estimates can give conflicting conclusions. We give an example of this tension when looking at a new chemotherapy for multiple myeloma.

Example 1.1 (*Carfilzomib-based Combination Therapy for Newly Diagnosed Multiple Myeloma (NDMM)*)**.** Until 2020, the effect of Carfilzomib-based combination therapy in the NDMM subpopu-

*Equal contribution. Order determined alphabetically. For code and instructions on data access, please visit: https://github.com/clinicalml/rct-obs-extrapolation

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

lation had not been studied via an RCT. However, a trial (ASPIRE) in 2015 measured the effect of Carfilzomib-based therapy on survival in Relapsed & Refractory Multiple Myeloma (RRMM) patients [58]. The CoMMpass trial, an observational dataset, was also available in which the Carfilzomib regimen was given to both NDMM and RRMM patients [36]. Several analyses on the CoMMpass dataset to estimate the effect of Carfilzomib-based therapy on NDMM patients led to different, sometimes opposing, conclusions on the benefit of the therapy in this subpopulation [33, 32].

A traditional meta-analysis approach would combine observational estimates under the assumption that differences arise only due to random variation, and not e.g., differences in confounding bias [24, Section 10.10.4.1]. This is unlikely to be true in practice. For instance, in Example 1.1, the two studies in question made different choices in e.g., how to adjust for confounders. *In this paper, we relax the assumption that all observational estimates are valid.* Instead, we assume that at least one observational estimate is valid across all subpopulations. In the context of Example 1.1, we might assume that at least one of the candidate observational studies yields consistent and asymptotically normal estimates of the effects in both the NDMM and RRMM populations. While we cannot *verify* that any given estimator is valid for all subpopulations, we can *falsify* this claim of validity if an estimator is inconsistent for the causal effects identified by the RCT (e.g., RRMM). Hence, we use the term *validation effects* to refer to causal effects in subpopulations that overlap between the observational and randomized datasets (e.g., RRMM), and use the term *extrapolated effects* to refer to those only covered by observational datasets (e.g., NDMM).

We propose a meta-algorithm that combines two key ideas: falsification of estimators, and pessimistic combination of confidence intervals. We first aim to falsify candidate estimators using hypothesis testing, rejecting those that fail to replicate the RCT estimates of validation effects. In Section 2.2, we motivate this approach with examples of observational estimates based on different causal assumptions, showing that hypothesis tests based on asymptotic normality can be applied even when causal assumptions fail to hold. Then, we combine accepted estimators to get confidence intervals on the extrapolated effects. Since failure to reject does not imply validity,[2] we return an interval that contains every confidence interval of the accepted estimators. We demonstrate theoretically that if *at least one* candidate estimator is consistent for both the validation and extrapolated effects, then the intervals returned by our algorithm provide valid asymptotic coverage of the true effects.

In scenarios where the covariate distribution differs across datasets, estimators that "transport" the causal effect should be used [42, 14, 15]. Furthermore, in the case of high-dimensional covariates, flexible machine learning methods are required to estimate nuisance functions, which can affect the hypothesis tests due to their slower convergence rates. In light of this, we adapt estimators of the average treatment effect in this setting to provide estimates of group-wise treatment effects, and show (via the framework of double machine learning [12, 55]) that this estimator enjoys asymptotic normality under mild conditions on convergence rates of the nuisance function estimators. Our conclusions are supported by semi-synthetic experiments, based on the IHDP dataset, as well as real-world experiments, based on clinical trial and observational data from the Women's Health Initiative (WHI), that demonstrate various characteristics of our meta-algorithm.

2 Setup and Motivating Examples

2.1 Notation and Assumptions

Let $Y \in \mathcal{Y}$ denote an outcome of interest, and $A \in \{0, 1\}$ denote a binary treatment. We use Y_a to denote the potential outcome of an individual under treatment $A = a$. We use $X \in \mathcal{X}$ to denote all other covariates. To distinguish between different sampling distributions (i.e., datasets), we use the random variable $D \in \{0, \dots J\}$, where $J \geq 1$ is the number of observational datasets, and $D = 0$ is reserved for the sampling distribution of the randomized trial. We let $\mathbb{P}(Y_1, Y_0, Y, A, X, D)$ denote the joint distribution over all variables, including unobserved potential outcomes. For instance, $\mathbb{P}(Y_1, Y_0, X \mid D = 0)$ denotes the distribution of potential outcomes and covariates in the RCT.

We seek to estimate conditional average treatment effects for a finite set of I subgroups $\{\mathcal{G}_i\}_{i=1}^{I}$. We assume subgroups are defined a-priori by a function $G : \mathcal{X} \mapsto \{1, \dots, I\}$, such that $G = i$ indicates that $X \in \mathcal{G}_i$. We use observational data precisely because not all groups are supported on the RCT

[2] For instance, we could fail to reject due to low power, or because falsification is impossible, due to differences in causal structure across subpopulations, as discussed in Appendix A.

dataset. To this end, we use $\mathcal{I}_R = \{i : \mathbb{P}(G = i \mid D = 0) > 0\}$ to denote the set of subgroups supported on the RCT dataset, and we let $\mathcal{I}_O$ denote the complement $\{1, \ldots, I\} \setminus \mathcal{I}_R$. We use $|\mathcal{I}_R|$ to denote the cardinality of a set, and assume that every observational dataset has support for all groups.

Assumption 2.1 (Support)**.** We assume that $\mathbb{P}(G = i, D = j) > 0$ for all $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$, i.e., all observational datasets ($D \geq 1$) have support for all groups.

Definition 2.1 (Validation and Extrapolated Effects)**.** We define the group average treatment effect (GATE)[3] as

$$\tau_i := \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases} \tag{1}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect.

Here, we focus on discrete subgroups, in part to reflect the practical reality of comparing RCTs to observational studies, where we may have large observational datasets with rich covariates but only have access to the published results of the RCT, which often provides estimates (with confidence intervals) for subgroup effects but not the raw data itself [56, Figure 4, for example]. In Def. 2.1, we allow for the fact that different datasets may have different distributions of effect modifiers. To have a well-defined effect of interest, we have chosen the reference dataset $D = 1$ arbitrarily, but in principle we could choose any of the observational datasets. We discuss further nuances of this definition under Assumption 2.3. By Def. 2.1, we often write these effects as a vector $\tau \in \mathbb{R}^I$. We use $\hat{\tau}(k) \in \mathbb{R}^I$ to denote an estimator, where $k \in \{0, \ldots, K\}$, with $\hat{\tau}(0)$ reserved to denote the estimator derived from the RCT data. The remainder are observational estimators.[4] In general, we use "hat" notation to refer to estimators, and refer to their population quantities without a hat. We use N_k to denote the number of samples used by each estimator. Throughout, we will assume that the RCT estimator is consistent.

Assumption 2.2. The RCT estimator $\hat{\tau}(0)$ is a consistent estimator of the (supported) dimensions of τ, such that for each $i \in \mathcal{I}_R$, $\hat{\tau}_i(0)$ is consistent for τ_i.

Below, our central assumption states that at least one observational estimator also enjoys consistency. We discuss examples of specific observational estimators in Section 2.2.

Assumption 2.3. There exists at least one observational estimator $\hat{\tau}(k) \in \mathbb{R}^I$, $k \geq 1$ that is a consistent estimator of $\tau \in \mathbb{R}^I$, such that for each $i \in \{1, \ldots, I\}$, $\hat{\tau}_i(k)$ is consistent for τ_i.

Remark 2.1. Assumption 2.3 is our primary non-trivial assumption, and in Appendix B, we give one example of causal assumptions (for a given observational study) under which the entire GATE vector τ is **identifiable** from observational data, and give an **estimator** of the resulting observational quantity which is asymptotically normal [41, 42, 40, 15, 16]. In order to compare observational estimates with experimental ones, Assumption 2.3 requires not only that the observational data is free of confounding, but also that the causal effect can be transported to the RCT population. This can be done so long as relevant effect modifiers are observed in both the RCT and observational study, but the latter requirement is satisfied automatically (without requiring RCT data) if e.g., treatment effects are constant within each subgroup G, or if the distribution of effect modifiers is the same between the RCT and observational study, in which case $\mathbb{E}[Y_1 - Y_0 \mid D, G] = \mathbb{E}[Y_1 - Y_0 \mid G]$. This represents one (conservative) failure mode of our approach, in which we may reject an observational estimator due to failures in transportability, even if it yields unbiased estimates of the extrapolated effects.

Assumptions 2.2 and 2.3 imply that there exists an observational estimator $\hat{\tau}(k)$ such that both $\hat{\tau}_i(k)$ and the RCT estimate $\hat{\tau}_i(0)$ are both consistent for the validation effects τ_i, $\forall i \in \mathcal{I}_R$. To validate this implication in finite samples, we will construct a statistical test to compare $\hat{\tau}_i(k)$ and $\hat{\tau}_i(0)$. Our general approach could be modified to use any valid test, but to facilitate further analysis, as well as explicit construction of confidence intervals, we additionally assume the following:

Assumption 2.4. All GATE estimators are pointwise[5] asymptotically normally distributed.

$$\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))/\hat{\sigma}_i(k) \xrightarrow{d} \mathcal{N}(0, 1) \tag{2}$$

for all $k \in \{0, , ..., K\}$, and for all i in $\mathcal{I}_R$ if $k = 0$ (the RCT estimator), and otherwise for all $i \in \mathcal{I}_R \cup \mathcal{I}_O$. Here, $\xrightarrow{d}$ denotes convergence in distribution, and $\hat{\sigma}_i^2(k)$ is an estimate of the variance that converges in probability to $\sigma_i^2(k)$, the asymptotic variance of $\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))$.

[3] We use this term in line with the literature [11, 28, 38, 55] and to distinguish it from the CATE function.

[4] We define $\hat{\tau}(0)$ as a vector in $\mathbb{R}^I$ for simplicity of notation, allowing the entries $\hat{\tau}_i(0), i \in \mathcal{I}_O$ to be arbitrary.

[5] Here, "pointwise" refers to the fact that each subgroup effect estimate is asymptotically normal.

Assumption 2.4 requires each estimator $\hat{\tau}(k)$ to be consistent and asymptotically normal for some $\tau(k)$, which may **not** be equal to τ. This is not a particularly strong assumption, as we discuss below.

2.2 Asymptotic Normality of Biased Estimators

In this section, we give two simple examples to illustrate the principle that multiple estimators $\hat{\tau}(k)$ may be asymptotically normal, even if they are asymptotically biased (i.e., $\tau(k) \neq \tau$). In both cases, there is a distinction between the *statistical* assumptions required to obtain asymptotic normality, and the *causal* assumptions required for $\tau(k)$ to identify the causal effect τ. For simplicity in both examples, we restrict to the setting of comparing one-dimensional estimates $\tau(k) \in \mathbb{R}$, which estimate the GATE, τ, in a single group $G = 1$ covered by all datasets. The statistical claims here also extend to GATE estimation with multiple groups [55].

Example 2.1 (Variation in confounding across datasets)**.** Suppose that there is one estimator of the GATE per observational dataset, and each estimator seeks to estimate the population quantity, $\tau(k) = \mathbb{E}[g_k(1, X_k) - g_k(0, X_k) \mid G = 1, D = k]$, where X_k denotes the controls used in each study, and $g_k(A, X_k) := \mathbb{E}[Y \mid A, X_k, D = k]$ and $m_k(X_k) := \mathbb{P}(A = 1 \mid X_k, D = k)$. We assume that $\eta < m_k(x) < 1 - \eta$ for some $\eta > 0$ for all x, k. Note that $\tau(k)$ is only a *statistical* quantity: identifying this with the *causal* quantity (the GATE) requires additional assumptions like unconfoundedness, that $Y_a \perp\!\!\!\perp A \mid X_k$ for the given dataset D. This assumption may hold for some datasets, but not others, particularly if the set of observed confounders X_k differs across datasets.

Regardless of the interpretation of $\tau(k)$, one can construct estimators of it that are consistent and asymptotically normal using flexible machine learning estimators.[6] One approach, given in Chernozhukov et al. [12], is to use double machine learning (DML), which employs cross-fitting to produce estimates $\hat{\tau}(k)$ based on the doubly-robust score [45], while using plug-in estimates $\hat{g}_k, \hat{m}_k$ based on machine learning models. This approach achieves asymptotic normality, $\sqrt{N_k}(\hat{\tau}(k) - \tau(k))/\hat{\sigma}^2(k) \xrightarrow{d} \mathcal{N}(0, 1)$, under regularity conditions that allow for flexible machine learning estimators that converge at slower than parametric rates, and where $\hat{\sigma}^2(k)$ converges in probability to the variance of the doubly robust score [See Theorem 5.1 of 12, for additional details]. These results hold whether or not $\tau(k) = \tau$, as discussed in Footnote 9 of Chernozhukov et al. [12]. For simplicity, we have focused on the case where $\mathbb{E}[Y_1 - Y_0 \mid G = 1]$ is constant across datasets. When this does not hold, certain conditions enable valid transportation of treatment effects across datasets [16] with the use of transported estimators [15] (see Appendix B for details).

Example 2.2 (Selection of Adjustment Strategy)**.** Consider the two causal graphs given in Figure 1, and assume that all variables are binary. Each graph suggests a different identification strategy for the causal effect, $\mathbb{E}[Y \mid do(A = a), G = 1]$. In Figure 1a, this is identified by the (observational) quantity $\mathbb{E}[Y \mid A = a, G = 1]$, and in Figure 1b, by front-door adjustment [39] as $\sum_M P(M \mid a, G = 1) \sum_{A'} \mathbb{P}(Y \mid M, A', G = 1)\mathbb{P}(A' \mid G = 1)$.

These observational quantities will typically differ: the one that represents the true interventional effect depends on which graph reflects the true causal structure. However, in the case where all variables are discrete and low-dimensional, we can still construct asymptotically normal estimators for both observational quantities.[7] For more complex settings (e.g., requiring regularized ML models for estimating conditional distributions) asymptotic normality has been established under certain conditions for general graphs [7, 29]

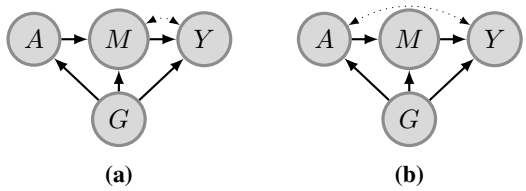


Figure 1: (Ex. 2.2) In (a), M and Y are confounded by unobservables (bi-directional dotted arrow). In (b), A and Y are confounded, but the causal effect is identified via front-door adjustment.

Remark 2.2. In each example, there are multiple estimators available, each asymptotically normal under basic statistical assumptions, but potentially biased in the sense that $\tau(k) \neq \tau$. In the first example, this bias occurs if X is not sufficient to control for confounding in all observational

[6] A rich literature focuses on establishing such results, beyond the approach in this example [2, 20, 60, 37, 3].

[7] This follows from the use of maximum likelihood (i.e., empirical counts) for estimating each conditional distribution, and applying the delta method to the front-door estimator.

datasets. In the second, this bias arises in a given estimator if the causal graph is incorrectly specified. Assumption 2.3 corresponds to assuming that both the statistical assumptions and causal assumptions hold for one of the candidate estimators, e.g., X is sufficient to control for confounding in at least one study (Example 2.1), or that one of the causal graphs is correct (Example 2.2).

2.3 Asymptotic Normality of GATE Estimators with Transportation

Example 2.1 assumes that $\mathbb{E}[Y_1 - Y_0 \mid G = i]$ is constant across datasets. In practice, it may be necessary to correct for differences (not captured by group indicators) between the observational and RCT populations. There exist estimators for the ATE in this setting under mild additional assumptions [15, 14]. These extend in a straightforward way to estimators of the GATE, but proving asymptotic normality is nuanced in high-dimensional settings when using flexible machine learning methods to estimate nuisance functions. For completeness, inspired by Semenova and Chernozhukov [55], we demonstrate that a doubly-robust GATE estimator for this setting is asymptotically normal under reasonable conditions (Assumption C.1 to C.5). Details on the estimator, and the corresponding proof of normality, are given in Appendix C, and may be of independent interest.

2.4 Testing for Bias under Asymptotic Normality

Under Assumption 2.4, each observational estimate $\hat{\tau}_i(k)$ can be compared to the estimate from the randomized trial $\hat{\tau}_i(0)$ for $i \in \mathcal{I}_R$, the groups with common support. Since the observational and randomized datasets are distinct, we can conclude that each $\hat{\tau}_i(k)$ is independent of $\hat{\tau}_i(0)$, and use this to test for the hypothesis that $\tau_i(k) = \tau_i$.

Proposition 2.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 2.2 and 2.4 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1-\rho)N$ for $\rho \in (0,1)$. Define the test statistic*

$$\hat{T}_N(k,i) := \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{3}$$

where $\hat{s}^2 := \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) := \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k,i) \xrightarrow{d} \mathcal{N}(0,1)$.

We present the proof for Proposition 2.1 in Appendix D. This asymptotic normality allows for the construction of simple hypothesis tests. For instance, one can construct a Wald test for $H_0 : \tau_i(k) = \tau_i$, with asymptotic level α by setting $\mu_i(k) = 0$ in Equation (3) and rejecting H_0 whenever, $|\hat{T}_N(k,i)| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the normal CDF. Moreover, the asymptotic power of this test (the probability of correctly rejecting H_0) is given by

$$1 - \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} + z_{\alpha/2}\right) + \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} - z_{\alpha/2}\right) \tag{4}$$

where $\sigma_{k,0}^2 := \frac{\sigma^2(k)_i}{N_k} + \frac{\sigma_i^2(0)}{N_0}$ [see Theorems 10.4, 10.6 of 62]. Likewise, Assumption 2.4 implies an asymptotic $1-\alpha$ confidence interval for $\tau_i(k)$ as

$$[\hat{L}_i(k)(\alpha), \hat{U}_i(k)(\alpha)] := \left[\hat{\tau}_i(k) - \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}, \hat{\tau}_i(k) + \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}\right] \tag{5}$$

3 Meta-Algorithm for Conservative Extrapolation

In this section, we more formally introduce our algorithm (Algorithm 1). There are two primary steps: falsification of estimators, and combination of confidence intervals. First, we attempt to falsify candidate estimators via hypothesis testing, rejecting estimator k whenever we are able to reject the null hypothesis $H_0 : \tau_i(k) = \tau_i, \forall i \in \mathcal{I}_R$. We use Bonferroni correction to control the false positive rate of the test. For the combination of confidence intervals, while we are unlikely to reject the "correct" estimator if one exists (Assumption 2.3), we may be unable to reject all "incorrect" (i.e., biased) estimators. This motivates the combination of confidence intervals (for the extrapolated effects) of the accepted estimators by taking the maximum and minimum bounds over all such intervals. Our main result characterizes the properties of our procedure, with proof in Appendix D.

Algorithm 1 Extrapolated Pessimistic Confidence Sets

Input: Desired coverage $1-\alpha$. For each $i \in \mathcal{I}_R$, RCT estimate $\hat{\tau}_i(0)$ and variance $\hat{\sigma}_i^2(0)$. For each $i \in \mathcal{I}_R \cup \mathcal{I}_O$, K candidate estimators $\hat{\tau}_i(k)$ and variances $\hat{\sigma}_i^2(k)$. Sample sizes $N_0, \ldots, N_K$.
Initialize: Empty candidate set $\hat{\mathcal{C}} \leftarrow \varnothing$
for $k = 1$ **to** K **do**
  Compute $\hat{T}_N(k,i), \forall i \in \mathcal{I}_R$, with $\mu_i(k) = 0$ (Eq. 3)
  if $\forall i \in \mathcal{I}_R, \left|\hat{T}_N(k,i)\right| \leq z_{\alpha/4|\mathcal{I}_R|}$, **then** $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} \cup \{k\}$
end for
for $i \in \mathcal{I}_O$ **do**
  $\hat{L}_i \leftarrow \min_{k \in \hat{\mathcal{C}}} \hat{L}_i(k)(\alpha/2)$ and $\hat{U}_i \leftarrow \max_{k \in \hat{\mathcal{C}}} \hat{U}_i(k)(\alpha/2)$ (Eq. 5)
end for
Return: $\hat{L}_i, \hat{U}_i$ for each $i \in \mathcal{I}_O$.

Theorem 3.1 (Properties of Algorithm 1)**.** *Under Assumptions 2.1 and 2.2, the output of Algorithm 1 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. *Under Assumptions 2.3 and 2.4, for each $i \in \mathcal{I}_O$,*

$$\lim_{N \to \infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{6}$$

2. *Under Assumption 2.4, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,*

$$\lim_{N \to \infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{7}$$

The first point says that for each extrapolated effect τ_i, the coverage of the final confidence interval $[\hat{L}_i, \hat{U}_i]$ is at least $1-\alpha$ in the limit. It follows from Assumption 2.3 and 2.4 that at least one estimator provides intervals $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ that achieve asymptotic coverage of $1-\alpha/2$. The result follows from our choice of threshold for the significance test as well as application of union bounds. The second point says that we will reject estimators that are not consistent for the validation effects, in the limit. Assumption 2.4 ensures that Proposition 2.1 holds for all estimators, so that this rejection is a consequence of the asymptotic power in Equation (4), going to 1 for a fixed bias as $N \to \infty$.

Remark 3.1. Equations (4) and (5) are useful for building further intuition. All of the candidate confidence intervals shrink at a rate of $O(1/\sqrt{N})$ as the overall sample size increases. For sufficiently large N, the width of our generated intervals will depend largely on our power to reject biased estimators, which will be higher for observational estimates with larger biases for validation effects.

4 Semi-Synthetic Experiments

4.1 Setup of Simulation

We generate semi-synthetic RCTs and observational datasets with covariates from the Infant Health and Development Program (IHDP), a randomized experiment on premature infants assessing the effect of home visits from a trained provider on the future cognitive performance [8]. The outcomes are simulated. Our data generation is based on the partial IHDP dataset used in [25], which includes $n_0 = 985$ observation, 28 covariates, and a binary treatment variable. We construct a scenario where there are four subgroups, defined by the infant's birth weight and maternal marital status: (high [$\geq$ 2000g], married), (low [$<$ 2000g], married), (high, single) and (low, single), which we shorthand as HM, LM, HS and LS. We include all subgroups in the observational studies, but exclude the latter two subgroups for the simulated RCT (i.e. only infants with married mothers are in the RCT).

For each simulated dataset, we generate 1 RCT and K observational studies. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = r \cdot n_0$. We performed weighted sampling to induce a different covariate distribution for observational studies, such that male infants, infants whose mothers smoked, and infants whose mothers worked during pregnancy are less prevalent. Then, we introduce confounding in the observational data, generating m_c continuous confounders and m_b binary confounders. Finally, we simulate outcomes in each

dataset, modifying the response surface given in Hill [25]. In our experiments, we may choose to conceal some confounders in each observational study to mimic unobserved confounding, denoting the number of concealed variables across the K studies as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$. For further details on confounder generation, outcome simulation, and confounder concealment, see Appendix F. Data generation parameters include K, r, m_c, m_b, $\mathbf{c_z}$, and the significance level α. By default, we set $K = 5$, $r = 10$, $m_c = 4$, $m_b = 3$, $\mathbf{c_z} = (0, 0, 2, 4, 6)$, and α = 0.05. The full hyperparameter search is provided in Appendix F, and details of hyperparameter tuning can be found in Appendix C.

4.2 Implementation and Evaluation of Meta-Algorithm

To implement Algorithm 1, we first obtain GATE estimates for the four subgroups and their estimated variances in each observational study, combining techniques from the DML and trasportability literature [55, 15]. Estimation details are shown in Appendices B and C. For the RCT, we stratify the data into the subgroups HM, LM and estimate the GATEs as the difference of mean outcomes between the treated and untreated. The z tests in Algorithm 1 are applied to both GATE estimates in the HM and LM subgroups ($|\mathcal{I}_R|= 2$), and the significance level of the tests is set at $\alpha/4$.

We evaluate performance using two main metrics: (1) the coverage probability of the output confidence intervals (ideally at least $1-\alpha$), and (2) the width of the confidence intervals (narrower is better). In addition to assessing the intervals produced by Algorithm 1, which we call *Extrapolated Pessimistic Confidence Sets (ExPCS)*, we will evaluate intervals produced by a variant of our algorithm, called *Extrapolated Optimistic Confidence Sets (ExOCS)*. In *ExOCS*, after falsifying estimators, we combine confidence intervals using a random-effects meta-analysis on the non-falsified observational studies. We compare *ExPCS* and *ExOCS* against two baselines. *Meta-Analysis* is a random-effects meta-analysis on all observational studies, as described in Section 6, with heterogeneity variance estimated via the DerSimonian-Laird moment method [17]. This baseline is the current standard for aggregating observational study results. The second baseline, *Simple Union*, uses the maximum upper bound and minimum lower bound of the $1 - \alpha$ confidence intervals across all observational studies, with no falsification procedure.[8]

4.3 Results

We perform three semi-synthetic experiments to assess the performance of our proposed meta-algorithm under different scenarios. The first experiment applies our algorithm under the default settings given in Section 4.1. In the second experiment, we vary the sample size ratio between the observational studies and the original RCT, r, from 1 to 10. In the third experiment, we vary the proportion of biased observational studies by setting $\mathbf{c_z}$ to be $(0, 0, 0, 0, 0)$, $(0, 0, 0, 0, 3)$, $(0, 0, 0, 3, 3)$ or $(0, 3, 3, 3, 3)$, corresponding to $0, 1, 2, 4$ studies being biased out of a total of 5 observational studies. Results for the latter two experiments are shown over 100 simulations of the datasets. Results for all experiments are shown in Figures 2, 3, and Figure 5 in Appendix G, respectively. We observe the following:

Meta-algorithm produces confidence intervals that cover the true GATE with nominal probability: We demonstrate in Figure 2 the application of our meta-algorithm (*ExPCS*), a variant of it (*ExOCS*), and two other baselines on one dataset. Our goal is to produce narrow confidence intervals that still cover the true GATEs in the extrapolated subgroups. The confidence intervals of *ExPCS* cover the true GATEs in the extrapolated subgroups with reasonable widths. In contrast, intervals produced by *Meta-Analysis* fail to cover the true GATE in both extrapolated subgroups due to the false assumption of unbiasedness across all studies. The *ExOCS* approach produces narrow intervals for the extrapolated effects, though it barely covers the true effect in the HS subgroup. This hints at the need for a conservative combination of non-falsified studies. However, an overly conservative approach (e.g. *Simple Union*) produces wide intervals that may be of little use for meaningful inference.

Although *Meta-Analysis* produces confidence intervals with inadequate coverage, its intervals for the married subgroups still have considerable overlap with the intervals produced by the RCT. This suggests that testing the meta-analyzed GATE estimates against the RCT GATE estimates may not be enough to demonstrate their validity. Compared to our *ExPCS* intervals, the lower bounds of the

[8]Note that *Simple Union* combines $1 - \alpha$ confidence intervals, while our approach combines $1 - \alpha/2$ confidence intervals to account for the probability of rejecting the "correct" estimator, if one exists. As a result, *Simple Union* intervals do not always strictly cover the intervals produced by *ExPCS*.

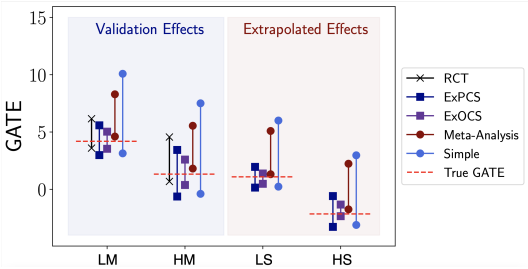
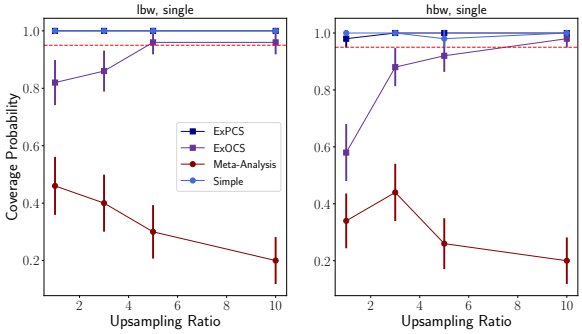


Figure 2: The confidence intervals for group average treatment effects (GATE) within the four subgroups output by our algorithm (*ExPCS*), our algorithm variant (*ExOCS*), random-effects meta-analysis on all observational studies (*Meta-Analysis*), simple union bound on all observational studies (*Simple*), and RCT, for one dataset generated using the default parameter settings laid out in Section 4.1. *LM, HM, LS, HS* represent four subgroups defined in Section 4.1



Mean width of 95% confidence intervals					
		Upsampling ratio			
Group	Estimator	1	3	5	10
LS	*ExPCS*	21.00	9.51	5.53	3.58
	Simple	21.00	10.20	6.95	5.69
HS	*ExPCS*	25.00	8.29	5.36	4.34
	Simple	26.00	9.15	6.90	6.33

Figure 3: Coverage probabilities of confidence intervals shown as a function of the size of the observational studies relative to the RCT. Dotted red lines stand for 95% coverage. Vertical bars are the 95% confidence intervals of the coverage probabilities. LS / HS stand for groups with low / high birth weight and single mother. Between *ExPCS* and *Simple* which have adequate coverage, *ExPCS* generally has narrower intervals.

Simple Union intervals are higher in several subgroups, since we use a higher confidence level for the candidate intervals corresponding to each study to account for probable error in study falsification.

An analysis of increasing observational study size: In Figure 3, we find that the coverage of the *Meta-Analysis* intervals is quite low across all sample sizes and particularly decreases at higher sample sizes. This result is intuitive, as three out of five studies are biased, meaning that meta-analysis will converge to a biased estimate as the amount of data increases. One could attempt to fix this issue through *ExOCS*, which does meta-analysis after falsification. However, *ExOCS* has poor coverage when the sample size of the observational studies is small, since the falsification tests are underpowered (evidenced by the high probability of selecting biased studies in Appendix G, Table 2). Both *ExPCS* and the *Simple Union* intervals have adequate coverage across all sample sizes. However, the widths of the intervals reported at the bottom of Figure 3 show that *ExPCS* intervals are narrower when there is adequate power, i.e. at higher sample sizes. Ultimately, *ExPCS* will tend to provide intervals that cover the true effect regardless of sample size, and in the case we have sufficient power, these intervals will both have good coverage and narrower width, allowing for more meaningful inference.

5 Women's Health Initiative (WHI) Experiments

In order to assess our approach in a real-world setting, we use clinical trial and observational data available from the WHI. Each subgroup is supported in both RCT and observational data, which proves useful for evaluation. At a high level, we "hide" some number of subgroups from the RCT, estimate a confidence interval of the effect estimate using our algorithm on the remaining data, and compare the result to the hidden RCT estimate. We do this over a large set of possible "held-out" subgroups, yielding >2000 different scenarios on which to test our approach. Because the original observational datasets replicate the RCT results fairly well using standard methods, we create additional "biased" datasets by sub-selecting the original observational dataset in a way that induces

selection bias. We evaluate each method, for each held-out subgroup, according to the length of the intervals as well as coverage of the RCT point estimates. Below, we describe the specifics of the data, the experimental setup, and the main results of the analysis. For additional details on data preprocessing, setup, and evaluation, see Appendix E.

5.1 Setup

The Postmenopausal Hormone Therapy (PHT) trial, i.e. the RCT used in this analysis, was run on postmenopausal women aged 50-79 years who had an intact uterus. It studied the effect of hormone combination therapy on several types of cancers, cardiovascular events, and fractures. The observational study (OS) was run in parallel, had a similar follow-up time to the RCT, and tracked similar outcomes. In our analysis, we use a composite outcome, where $Y = 1$ if any of the following events are **observed** to occur in the first 7 years of follow-up, and $Y = 0$ otherwise: coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. This represents a binarization of the "global index" time-to-event outcome from the original study, where $Y = 0$ could also occur due to censoring. We establish treatment and control groups in the OS based on explicit confirmation or denial of usage of both estrogen and progesterone in the first three years. We use only covariates measured in both the RCT and OS to simplify analysis.

5.2 Evaluation

Our empirical evaluation consists of several steps. In the first step, we replicate the principal results from the PHT trial, given in Table 2 of [51], by fitting a doubly robust estimator (of the style given in Appendix C) on the WHI OS data. Then, while treating the WHI OS dataset as the "unbiased" observational dataset, we simulate additional "biased" observational datasets by inducing selection bias into the WHI OS. The exact mechanism of selection bias and its clinical intuition is given in Appendix E. Importantly, this is the only part of the evaluation that involves any simulation.

The second step is to construct a large suite of tasks on which to evaluate our method, by considering different sets of validation-extrapolation subgroups. To construct the subgroups, we consider all pairs of a selected set of binary covariates (see Appendix E.6), where each pair defines four subgroups. For example, one covariate pair is ("current smoker", "currently drinks alcohol"). We treat two of the subgroups as validation subgroups and two as extrapolated subgroups. For the latter groups, we apply our algorithm without access to the RCT data, and only use the RCT data for final evaluation. The total number of covariate pairs is 592, leading to 1184 distinct "tasks" (i.e., extrapolated groups). For each task, we evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis (described in Section 4.2). Additionally, we evaluate an "oracle" method, which is identical to ExPCS, except that it always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias). For each method, we compute the following metrics, averaged across all tasks – **Length**: length of the confidence interval, **Coverage**: percentage of tasks where the interval covers the RCT point estimate. In addition, we report the **Unbiased OS Percentage**: the percentage of tasks where the ExPCS approach retains the unbiased study after the falsification step.

5.3 Results

Table 1 reports the metrics above, averaged across all extrapolated subgroups.

Compared to the "simple" baseline, our approach has better coverage with much shorter confidence intervals. Our falsification procedure retains the unbiased observational study 99% of the time, yielding near-oracle coverage rates, but produces substantially shorter intervals than the "simple" baseline. Recall that the simple baseline takes a union over all $1 - \alpha$ intervals estimated from each observational dataset, while ExPCS takes a union of a smaller number of slightly wider $(1 - \alpha/2)$ confidence intervals.

	Coverage	**Length**	**OS** %
Simple	0.39	0.416	–
Meta-Analysis	0.03	0.260	–
ExOCS	0.28	0.058	–
ExPCS (ours)	0.45	0.081	0.99
Oracle	0.44	0.068	–

Table 1: Coverage, length, and unbiased OS % of ExPCS and baselines. ExPCS achieves comparable coverage to the oracle method with highly efficient intervals. Additionally, we do not reject the unbiased OS in 99% of the tasks.

Compared to the Meta-Analysis and ExOCS baselines, we achieve comparable (or much better) length with substantially better coverage. In particular, compared to meta-analysis, we achieve tighter intervals and also cover the RCT estimate with higher frequency. This result is intuitive, since one will get a biased estimate if biased observational studies are included in the meta-analysis. Additionally, conservatively combining the non-falsified estimates (as opposed to *ExOCS*, which does a meta-analysis on the non-falsified estimates) is important to achieve good coverage (0.45 vs 0.28).

We get comparable coverage and interval lengths to the oracle method. Our coverage rate is nearly identical (0.45) to that of the oracle method (0.44), with intervals that are marginally wider (0.081 vs. 0.068). Our slightly improved coverage is possible due to the wider intervals. Note that our measure of "coverage" may be pessimistic, because we track coverage of the RCT point estimate, as opposed to the true causal effect (which is unknown), and the confidence intervals are designed to cover the latter. Indeed, we report the oracle method precisely as a means of providing a more suitable comparison. Overall, our real-world results suggest that our method of falsification followed by a conservative combination of intervals may be useful for biostatisticians and clinicians when doing meta-analyses.

6 Related Work

Meta-analysis for combining observational estimates Among the quantitative approaches for meta-analysis to account for potential bias, our *Meta-Analysis* baseline is standard for meta-analysis of observational data [24] to account for heterogeneity. Allowing for heterogeneity of treatment effects among studies produces wider confidence intervals and thus more conservative inference. If additional study-level covariates are available (e.g. study designs, drop-out rate), several approaches aim to adjust for potential bias, either by modeling the bias magnitude [19, 64, 1, 21], down-weighting studies with higher risk of bias [26, 35], or using Bayesian hierarchical regression to account for difference between subgroups of studies [44, 63]. Our work differs from these approaches, in that (1) we use information from outside the population of interest to assess bias, and (2) we do not place any assumptions on the patterns of bias across studies.

Partial identification and sensitivity analysis These methods seek to place bounds on causal effects when they cannot be point-identified. Our method can be seen as an alternative way of doing so, with a fundamentally different type of assumption. Methods for partial identification rely on having discrete variables and a known causal graph (typically including unobserved confounders) [18, Section 9]. Methods for sensitivity analysis, on the other hand, translate assumptions about the strength and nature of unobserved confounding into bounds on causal effects [47, 48, 65]. In contrast, we do not make any such assumptions, e.g., we allow for continuous variables, and when some candidate estimators are biased due to unmeasured confounding, we do not place any limit a-priori on the bias. An extended related work is given in Appendix H.

7 Discussion and Limitations

We have presented a meta-algorithm that constructs conservative confidence intervals for group average treatment effects of subgroups that are not represented in RCTs, but are represented in observational studies. Under the assumption that there exists at least one candidate estimator that is asymptotically normal and consistent for both the validation and extrapolated effects, these intervals will achieve the correct asymptotic coverage of the true effect. However, our method is not without limitations. Most notably, we may fail to reject the null hypothesis due to low power, e.g., when an observational estimate $\hat{\tau}(k)$ has high variance. In practice, we expect that our approach will be most useful when the observational studies in question have large sample sizes, leading to higher-precision estimates of potential bias, and smaller confidence intervals on the extrapolated effects. Our hope is that methods such as ours will lead to higher confidence in observational estimates when RCT data is available to falsify observational studies that do not replicate known causal relationships. Finally, great care should be taken to appropriately validate and soundly interpret the results of our method in practice, especially with more sensitive subgroups (e.g. with respect to race or gender).

Acknowledgments: We would like to thank Ahmed Alaa, Hunter Lang, Christina Ji, Hussein Mozannar, and other members of the Clincal Machine Learning group for helpful discussions and valuable feedback on the manuscript. ZH was supported by an ASPIRE award from The Mark Foundation for Cancer Research and by the National Cancer Institute of the National Institutes of Health under Award Number F30CA268631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MO and DS were supported in part by Office of Naval Research Award No. N00014-21- 1-2807. MCS was supported by the LEAP program from the Ministry of Science and Technology in Taiwan.

References

[1] Andrew Anglemyer, Hacsi T Horvath, and Lisa Bero. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane database of systematic reviews*, (4):MR000034, April 2014.

[2] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-Biased inference of average treatment effects in high dimensions. *arXiv preprint (1604.07125)*, April 2016.

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

[4] Hailey R Banack, Jay S Kaufman, Jean Wactawski-Wende, Bruce R Troen, and Steven D Stovitz. Investigating and remediating selection bias in geriatrics research: the selection bias toolkit. *Journal of the American Geriatrics Society*, 67(9):1970–1976, 2019.

[5] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

[6] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[7] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint (2003.12659)*, March 2020.

[8] Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3): 350–359, 1992.

[9] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[10] Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

[11] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint (1712.04802)*, December 2017.

[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The econometrics journal*, 21(1):C1–C68, January 2018.

[13] Jordan M Cloyd, Victor Heh, Timothy M Pawlik, Aslam Ejaz, Mary Dillhoff, Allan Tsung, Terence Williams, Laith Abushahin, John FP Bridges, and Heena Santry. Neoadjuvant therapy for resectable and borderline resectable pancreatic cancer: a meta-analysis of randomized controlled trials. *Journal of clinical medicine*, 9(4):1129, 2020.

[14] Issa J Dahabreh, Sarah E Robertson, Lucia C Petito, Miguel A Hernán, and Jon A Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *arXiv preprint (1908.09230)*, August 2019.

[15] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, June 2020.

[16] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint (2102.11904)*, February 2021.

[17] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[18] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. September 2021.

[19] David M Eddy, Vic Hasselblad, and Ross Shachter. An introduction to a bayesian method for meta-analysis: the confidence profile method, 1990.

[20] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint (1309.4686)*, February 2018.

[21] Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.

[22] Gordon H Guyatt, Andrew D Oxman, Regina Kunz, Gunn E Vist, Yngve Falck-Ytter, and Holger J Schünemann. What is "quality of evidence" and why is it important to clinicians? *Bmj*, 336(7651):995–998, 2008.

[23] Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, 336(7650):924–926, 2008.

[24] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

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

[26] Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

[27] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, April 2015.

[28] Daniel Jacob. Group average treatment effects for observational studies. *arXiv preprint (1911.02688)*, November 2019.

[29] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *AAAI*. aaai.org, 2021.

[30] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett, editor, *Advances in Neural Information Processing Systems*, 31. Curran Associates, Inc., 2018.

[31] N Keum, DH Lee, DC Greenwood, JE Manson, and E Giovannucci. Vitamin d supplementation and total cancer incidence and mortality: a meta-analysis of randomized controlled trials. *Annals of Oncology*, 30(5):733–743, 2019.

[32] Ola Landgren, David S Siegel, Daniel Auclair, Ajai Chari, Michael Boedigheimer, Tim Welliver, Khalid Mezzi, Karim Iskander, and Andrzej Jakubowiak. Carfilzomib-lenalidomide-dexamethasone versus bortezomib-lenalidomide-dexamethasone in patients with newly diagnosed multiple myeloma: results from the prospective, longitudinal, observational commpass study. *Blood*, 132:799, 2018.

[33] Bin Li, Kaili Ren, Lei Shen, Peijie Hou, Zhenqiang Su, Alessandra Di Bacco, Jin-Liern Hong, Aaron Galaznik, Ajeeta B Dash, Victoria Crossland, et al. Comparing bortezomib-lenalidomide-dexamethasone (vrd) with carfilzomib-lenalidomide-dexamethasone (krd) in the patients with newly diagnosed multiple myeloma (ndmm) in two observational studies. *Blood*, 132:3298, 2018.

[34] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, May 2010.

[35] Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in medicine*, 28(28):3562–3566, 2009.

[36] NIH. Relating clinical outcomes in multiple myeloma to personal assessment of genetic profile (com-mpass). *Clinical Trials website. https://clinicaltrials. gov/ct2/show/NCT01454297*, 2016.

[37] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4932–4941. PMLR, 2019.

[38] Chan Park and Hyunseung Kang. A groupwise approach for inferring heterogeneous treatment effects in causal inference. *arXiv preprint (1908.04427)*, August 2019.

[39] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[40] Judea Pearl. Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266, September 2015.

[41] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254, August 2011.

[42] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, November 2014.

[43] Francesco Paolo Prete, Angela Pezzolla, Fernando Prete, Mario Testini, Rinaldo Marzaioli, Alberto Patriti, Rosa Maria Jimenez-Rodriguez, Angela Gurrado, and Giovanni FM Strippoli. Robotic versus laparoscopic minimally invasive surgery for rectal cancer: a systematic review and meta-analysis of randomized controlled trials. *Annals of surgery*, 267(6):1034–1046, 2018.

[44] Teresa C Prevost, Keith R Abrams, and David R Jones. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in medicine*, 19(24):3359–3376, 2000.

[45] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, March 1995.

[46] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[47] P R Rosenbaum and D B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 45(2):212–218, 1983.

[48] Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

[49] Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*, 2020.

[50] Evan TR Rosenman, Art B Owen, Mike Baiocchi, and Hailey R Banack. Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 2021.

[51] Jacques E Rossouw, Garnet L Anderson, Ross L Prentice, Andrea Z LaCroix, Charles Kooperberg, Marcia L Stefanick, Rebecca D Jackson, Shirley AA Beresford, Barbara V Howard, Karen C Johnson, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama*, 288(3):321–333, 2002.

[52] Peter F Schnatz, Xuezhi Jiang, Aaron K Aragaki, Matthew Nudy, David M O'Sullivan, Mark Williams, Erin S LeBlanc, Lisa W Martin, JoAnn E Manson, James M Shikany, et al. Effects of calcium, vitamin d, and hormone therapy on cardiovascular disease risk factors in the women's health initiative: a randomized controlled trial. *Obstetrics and gynecology*, 129(1):121, 2017.

[53] Martijn J Schuemie, Patrick B Ryan, William DuMouchel, Marc A Suchard, and David Madigan. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218, January 2014.

[54] Martijn J Schuemie, George Hripcsak, Patrick B Ryan, David Madigan, and Marc A Suchard. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2571–2577, March 2018.

[55] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The econometrics journal*, 24(2):264–289, 2021.

[56] SPRINT Research Group, Jackson T Wright, Jr, Jeff D Williamson, Paul K Whelton, Joni K Snyder, Kaycee M Sink, Michael V Rocco, David M Reboussin, Mahboob Rahman, Suzanne Oparil, Cora E Lewis, Paul L Kimmel, Karen C Johnson, David C Goff, Jr, Lawrence J Fine, Jeffrey A Cutler, William C Cushman, Alfred K Cheung, and Walter T Ambrosius. A randomized trial of intensive versus standard Blood-Pressure control. *The New England journal of medicine*, 373(22):2103–2116, November 2015.

[57] Leonard A Stefanski and Dennis D Boos. The calculus of M-Estimation. *The American statistician*, 56(1):29–38, 2002.

[58] A Keith Stewart, S Vincent Rajkumar, Meletios A Dimopoulos, Tamás Masszi, Ivan Špička, Albert Oriol, Roman Hájek, Laura Rosiñol, David S Siegel, Georgi G Mihaylov, et al. Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma. *New England Journal of Medicine*, 372(2):142–152, 2015.

[59] Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, Ruijun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, and Patrick B Ryan. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826, November 2019.

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

[61] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[62] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY, 2004.

[63] Nicky J Welton, Anthony E Ades, JB Carlin, DG Altman, and JAC Sterne. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):119–136, 2009.

[64] Robert L Wolpert and Kerrie L Mengersen. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*, 19(3):450–471, 2004.

[65] Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. August 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Notation and Assumptions, Section 2.1 and Appendices B and C.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

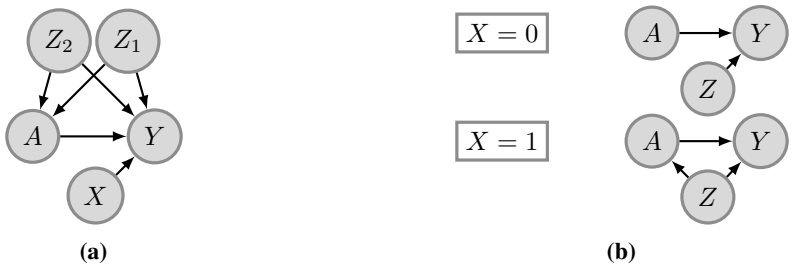


Figure 4: Example A.1 is depicted in (a), and Example A.2 in (b)

A When can biased estimators be falsified?

As discussed in Examples 2.1 and 2.2, we imagine that observational estimators differ in a few possible ways. They may represent the same identification strategy applied to different datasets, different identification strategies applied to the same dataset (e.g., different choices of confounders), or some combination of the two.

Assumption 2.3 states that there exists a consistent and asymptotically normal observational estimator for τ, as defined in Def. 1. This is a fundamental assumption in our work, and so we build additional intuition for when we might expect this condition to hold, and when we might be able to falsify this assumption. In this section, we give basic intuition regarding patterns of confounding, and in Section B, we discuss issues of transportability.

In Example A.1, we give a simple example where the causal graph is consistent across two subgroups, and where an estimator must control for all confounders to get consistent estimates of the GATE in either subgroup. In this setting, falsification is possible. On the other hand, in Example A.2, we give a counterexample, where there are multiple estimators that can deliver consistent estimates of the GATE on the RCT subpopulation, but only one provides consistent estimates across all subpopulations.

Example A.1 (Consistent confounding across subgroups)**.** In the causal graph shown in Figure 4a, there are two sets of confounders, Z_1, Z_2, a binary treatment variable A, a binary subgroup variable X, and the outcome Y. We assume a linear outcome model, whereby $E[Y|X, Z_1, Z_2, A] = \alpha + \beta X + \gamma_1 AX + \gamma_2 A(1 - X) + \delta_1 Z_1 + \delta_2 Z_2$. Note that the true group average treatment effect (GATE) for the two subgroups are, GATE$(X = 0) = \gamma_2$; GATE$(X = 1) = \gamma_1$. It is straightforward to show that not conditioning on the full set of confounders will lead to biased GATE estimates for both subgroups, whereas conditioning on both Z_1 and Z_2 will lead to consistent estimates for both subgroups.

Example A.2 (Selective confounding by subgroup)**.** Let there be two subgroups, $X = 0$ and $X = 1$, with the former having support in both RCT and observational studies and the latter having support in only observational data. Now, suppose we had the following treatment assignment mechanism, $p(A = 1|X, Z) = f(Z) \cdot \mathbf{1}(X = 1) + c \cdot \mathbf{1}(X = 0)$, where Z is a set of confounders, f is a nonlinear function of Z, and c is a constant. A candidate estimator that does not condition on Z would be able to get consistent estimates for the validation effect but not the extrapolated effect. On the other hand, conditioning on Z would allow for consistent estimates on both validation and extrapolated effects.

B Conditions for valid observational / randomized comparisons

Recall that we had defined the group average treatment effect (GATE) as follows in Equation (1)

$$\tau_i := \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases}, \tag{8}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect. In this section, we discuss sufficient conditions under which these causal effects are identifiable from observational data drawn from a distribution $D = k$, and give examples of doubly-robust estimators of these quantities. These assumptions cover both comparisons of the observational studies to the randomized trial (used for validation), as well as the normalization of observational estimates (used for confidence intervals on the extrapolated effects).

Our goal in presenting these results is to build intuition in this setting for when we might expect a consistent observational estimator to exist across all groups. This is a well-studied topic, often in the context of generalizing effect estimates from randomized trials to other supported populations (e.g., all trial-eligible individuals). We primarily make use of results in that literature to build intuition here, pointing the reader to Degtiar and Rose [16] for a recent review whose presentation we largely mirror, with modifications to account for our notation.

B.1 Identification

First, we state standard assumptions under which the GATE in the observational population for $D = k$,

$$\mathbb{E}[Y_1 - Y_0 \mid G = i, D = k], \tag{9}$$

is identifiable from data in the dataset $D = k$, with notation adapted to our setting.

Assumption B.1. The following conditions hold for the distribution $\mathbb{P}(\cdot \mid D = k)$:

1. *Conditional Exchangeability over A*: $Y_a \perp\!\!\!\perp A \mid X, D = k$ for all treatments a.
2. *Positivity of Treatment Assignment*: $\mathbb{P}(X = x \mid D = k) > 0 \implies \mathbb{P}(A = a \mid X = x, D = k) > 0$ for all a.
3. *Consistency*: $A = a \implies Y_a = Y$

These causal assumptions ensure that the ATE and CATE can be identified from observational data for the observational population and are standard in the causal inference literature [27]. In order to transport these estimates to the RCT population (or from one observational dataset to another), we require additional assumptions. Next, we give assumptions under which these estimates can be transported to another population $D = k'$, where in our case $k' \in \{0, 1\}$.

Assumption B.2. Let k correspond to a source population, and k' correspond to the target population. Conditioned on the event $D \in \{k, k'\}$, define the random variable $S = 1$ if $D = k$ and $S = 0$ otherwise. Then let the following hold, on the distribution $\mathbb{P}(\cdot \mid D \in \{k, k'\})$.

1. *Conditional Exchangeability over S*: $Y_a \perp\!\!\!\perp S \mid X$ for all treatments a.
2. *Positivity of Selection*: $\mathbb{P}(X = x) > 0 \implies \mathbb{P}(S = 1 \mid X = x) > 0$ almost surely over X for all a.
3. *Consistency*: $S = s$ and $A = a \implies Y_a = Y$

Here, we note that this introduces non-trivial additional assumptions. Most notably, we require that the potential outcomes are independent of the dataset, given X. This would be violated, for instance, if the distribution of unobservable effect modifiers differs between different observational studies. As a result, we note that it is possible for an observational study to fail to replicate the RCT results due to failures of transportability (failure of Assumption B.2) even if it has "internal validity", allowing for identification of the causal effect in the population $D = k$. There also exists a large body of work on identifying transportable causal effects via causal graphs [41, 42, 40].

B.2 Estimation of the ATE in the target population

Regarding estimation, Dahabreh et al. [15] consider the problem of transporting average treatment effects from randomized trials to observational studies, under Assumption B.1 with $k = 0$ and Assumption B.2 with $k = 0, k' = 1$. These assumptions admit identification of the potential outcomes means as follows (see Section 4.2 of Dahabreh et al. [15])

$$\mathbb{E}[Y_a \mid S = 0] = \mathbb{E}[\mathbb{E}[Y \mid X, S = 1, A = a] \mid S = 0] \tag{10}$$

where the outer expectations are over $\mathbb{P}(X \mid S = 0)$, i.e., the covariate distribution of the target population. Dahabreh et al. [14] give a doubly robust estimator for the statistical quantity on the right-hand side as the empirical expectation of the following pseudo-outcome (see Equation A.13 of Dahabreh et al. [15])

$$\hat{\mu}(a) = \frac{1}{n} \sum_{i=1}^{n} Y_i^a(\hat{\eta}, \hat{\pi}) \tag{11}$$

where n is the total samples in both the source $S = 1$ and target $S = 0$ samples, and where

$$Y_i^a(\hat{\eta}, \hat{\pi}) := \frac{1}{\hat{\pi}} \left(\mathbf{1}\{S_i = 1, A_i = a\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)\hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} + (1 - S_i)\hat{g}_a(X_i) \right). \quad (12)$$

In Equation (12), $\hat{\eta} := (\hat{g}_a, \hat{e}_a, \hat{p})$, and $\hat{\pi} := n^{-1} \sum_{i=1}^n \mathbf{1}\{S_i = 0\}$ is an estimate of $\mathbb{P}(S = 0)$, $\hat{g}_a(X)$ is an estimate of the mean conditional outcome $\mathbb{E}[Y \mid A = a, S = 1, X]$, $\hat{p}(X)$ is an estimate of the selection probability $\mathbb{P}(S = 1 \mid X)$, and $\hat{e}_a(X)$ is an estimate of the propensity score $\mathbb{P}(A = a \mid S = 1, X)$. Dahabreh et al. [14] derives precise asymptotic properties of this estimator, which is asymptotically normal and consistent for the observational quantity on the right-hand side of Equation (10). In particular, this estimator is doubly-robust in the sense that it is consistent if either $\hat{p}(X)$ or $\hat{g}_a(X)$ is consistent, but requires consistency of $\hat{e}_a(X)$. It also enjoys the rate double-robustness property, retaining consistency and asymptotic normality even if the estimators for $\hat{p}, \hat{g}$ converge at slower than parametric rates, and allows for the same cross-fitting schemes used in the Double ML [12] literature for relaxing Donsker conditions.

Note that the average treatment effect in this setting can be estimated by the following contrast, which is similarly an empirical expectation of a pseudo-outcome

$$\hat{\mu}(1) - \hat{\mu}(0) = \frac{1}{n} \sum_{i=1}^n Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi}) = \frac{1}{n} \sum_{i=1}^n Y_i(\hat{\eta}, \hat{\pi}), \quad (13)$$

where $Y_i(\hat{\eta}, \hat{\pi}) := Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi})$. Furthermore, the variance of these estimates can be estimated using either sandwich estimators from M-estimation theory [57], or via bootstrap methods. We refer the reader to Sections 5.3, 5.4 and Appendix A.4 of [15] for more details.

C Estimation and comparison of GATE in semi-synthetic experiments

In Sections 2.2 and B, we discuss several estimators for average treatment effects (ATEs) that are known to be asymptotically normal, such as the double ML estimator discussed in Example 2.1 or the doubly-robust estimator in Section B.

Given a fixed set of discrete subgroups, one could analyze each subgroup independently and apply such estimators directly, since the ATE in each subgroup is precisely the GATE. This would be a straightforward way to ensure that the same formal guarantees hold regarding asymptotic normality. While this approach would be feasible in our experimental setting, due to the small number of groups, it is less practical in general, especially with a larger number of groups, since information cannot be shared across nuisance models such as $\hat{g}_a, \hat{e}_a, \hat{p}$ discussed in Section B.

In an effort to emulate a more realistic setting, we take a slightly different approach in the semi-synthetic experiments. We draw inspiration from the double ML approach given in Semenova and Chernozhukov [55] for GATE estimation, while taking into consideration the transportation of causal effects in the sense of Section B. Note that in Semenova and Chernozhukov [55], the required assumptions and proofs for asymptotic normality of estimators are provided on a case-by-case basis, which does not include our case with transportation. Therefore, in the following we will briefly describe their approach, then show how we construct our GATE estimators and provide the required assumptions for their asymptotic normality.

Semenova and Chernozhukov [55] focuses on the setting where there exists some pseudo-outcome / signal, $Y(\eta)$, and where one is interested in summarizing the function, $\tau(x) = \mathbb{E}[Y(\eta) \mid X = x]$, with a linear regression function (in the simplest case, a set of group indicators). When $Y(\eta)$ is the doubly-robust score [46, 45] (see Equation (18)), $\tau(x)$ is equal to the CATE function, and the best approximation by group indicators gives the GATE.

Our general procedure is as follows: for estimation of $\hat{\tau}(k)$ and the respective variances, we construct a score function / pseudo-outcome, $\tilde{Y}$, whose empirical conditional expectation (in each group) provides an estimate of the GATE, and whose empirical variance we use as an estimate of the variance. We describe this procedure in more detail below. Throughout, X should be taken to refer to the covariates that are observed in a given observational study.

Comparing Validation Effect Estimates In our simulation setup, all of the observational datasets are drawn from a common distribution, which differs from the RCT distribution, requiring the use of

the techniques and assumptions discussed in Section B to estimate the GATE, $\tau_i = \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0]$, using data from the observational distributions.

To generate the observational estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ in this setting, we cannot simply take empirical conditional expectation / variance of the score function given in Equation 13. Rather, the GATE is identified under Assumptions B.1 and B.2 as a conditional expectation of the score times a correction factor, as discussed in the following proposition.

Proposition C.1. *In the setting of Section B, under Assumptions B.1 and B.2, the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S = 0, G = i] = \frac{\mathbb{P}(S = 0)}{\mathbb{P}(S = 0 \mid G = i)} \mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i], \tag{14}$$

where $Y^a(\eta, \pi)$ is defined as in Equation 15.

$$Y^a(\eta, \pi) := \frac{1}{\pi} \left(\mathbf{1}\{S = 1, A = a\} \cdot \frac{1 - p(X)}{p(X) e_a(X)} \cdot \{Y - g_a(X)\} + (1 - S) g_a(X) \right) \tag{15}$$

where $\eta := (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi := \mathbb{P}(S = 0)$ with true value π_0, $g_a(X) := \mathbb{E}[Y \mid A = a, S = 1, X]$, $p(X) := \mathbb{P}(S = 1 \mid X)$, and $e_a(X) := \mathbb{P}(A = a \mid S = 1, X]$.

A proof is provided in Appendix D. Note that this is equivalent to replacing the estimate of $1/\mathbb{P}(S = 0)$ in the score with an estimate of $1/\mathbb{P}(S = 0 \mid G = i)$, before computing the empirical conditional expectations of the score.

Now, for each observational dataset, we construct estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ for $i \in \mathcal{I}_R$ as follows:

1. We collect observational samples from the two validation groups {lbw, married} and {hbw, married}, which we denote as $G = 0, G = 1$ respectively. We combine these observational samples with the samples from the RCT, using $S = 0$ to denote RCT samples (the target distribution) and $S = 1$ to denote observational samples.
2. We define our signal for each sample as

$$\tilde{Y}_i(\hat{\eta}, \hat{\pi}_g) := \tilde{Y}_i^1(\hat{\eta}, \hat{\pi}_g) - \tilde{Y}_i^0(\hat{\eta}, \hat{\pi}_g) \tag{16}$$

   where we define the modified score $\tilde{Y}_i^a$, in light of Proposition C.1, as

$$\tilde{Y}_i^a(\hat{\eta}, \hat{\pi}_g) := \frac{1}{\hat{\pi}_g(G_i)} \left(\mathbf{1}\{S_i = 1, A_i = a\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i) \hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} + (1 - S_i) \hat{g}_a(X_i) \right), \tag{17}$$

   where $\hat{\pi}_g(G_i)$ is defined as an estimate of $\pi_g(G_i) := \mathbb{P}(S = 0 \mid G_i)$, computed using empirical averages.
3. We use 3-fold cross-fitting as described in Semenova and Chernozhukov [55] to generate the signals for each sample, such that for the i-th datapoint, the score $\tilde{Y}_i(\hat{\eta}, \hat{\pi})$ uses plug-in estimates $\hat{\eta} = (\hat{g}_1, \hat{g}_0, \hat{e}_1, \hat{p})$ that are learned on the folds that do not include the i-th datapoint, and $\hat{\pi}$ is estimated using empirical averages. In practice, we use a multi-layer perceptron (MLP) regressor for estimating $\hat{g}_a$, and ℓ_2-regularized logistic regression for estimating $\hat{e}_1, \hat{p}$, with hyperparameters described in Section F. For each model, we reserve 20% of the current fold in the cross fitting procedure as a validation set to do hyperparameter selection.
4. Finally, we estimate $\hat{\tau}_i(k)$ as the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g) \mid G = i]$, and we use the empirical conditional variance of this score to estimate the variance $\hat{\sigma}_i^2(k)$.

We construct the RCT estimate $\hat{\tau}_i(0)$ (using the RCT sample alone) as the difference of the empirical conditional means $\mathbb{E}_{N_0}[Y \left(\frac{\mathbf{1}\{A=1\}}{\hat{P}(A=1)} - \frac{\mathbf{1}\{A=0\}}{1-\hat{P}(A=1)} \right) \mid G = i]$, where $\hat{P}(A = 1)$ is an empirical average. We compute $\hat{\sigma}_i^2(0)$ as the empirical conditional variance of this quantity. We then conduct testing, as described in Algorithm 1.

Asymptotic normality of transported estimators We herein provide sufficient assumptions that guarantee the asymptotic normality of our transported GATE estimators, i.e. the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$:

Assumption C.1 (Observational dataset covers the whole support of covariates)**.**

$$\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0$$

Note that Assumption C.1 is implied by Assumption 2.1.

Assumption C.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x)|G = i, S = 0] = \sigma^2_{\tau i} < \infty$$

Assumption C.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption C.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2|X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

Assumption C.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers
- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)} \in \mathcal{T}_n) \geq 1 - \epsilon_n$
- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions g_1, g_0, e_1, p, defined as follows:

$$\mathbf{g}_n \coloneqq \max_{a \in \{0,1\}} \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2}$$
$$\mathbf{e}_n \coloneqq \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2}$$
$$\mathbf{p}_n \coloneqq \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta \in \cup_{n=1}^{\infty} \mathcal{T}_n} \left(\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} |g_a(x)| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{p(x)} \right| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{e_1(x)} \right| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{e_0(x)} \right| \right) = \bar{\mathcal{C}} < \infty$$

Theorem C.1. *Suppose Assumptions C.1 to C.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$, where $\tilde{Y}$ is defined in Equation 16 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Remark: As we will prove later in section D.2, Assumptions C.1 to C.4 guarantee that when the nuisance function vector $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is known (i.e. need not be estimated), the transported GATE estimator is asymptotically normal. In practice, $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is not known and has to be estimated, so Assumption C.5 lays out sufficient properties the nuisance function vector estimator

needs to satisfy. In particular, Assumptions C.5.A and C.5.B permit that the convergence rate of estimators can be slower than $o_P(n^{-1/2})$, which is useful when X is high-dimensional and machine learning models are required to estimate the nuisance functions. To date, a variety of commonly-used machine learning models have been shown to enjoy a convergence rate of at least $o_P(n^{-1/4})$, e.g. [9, 6, 5] for certain ℓ_1 penalized models, [61] for a class of regression trees and random forests, and [10] for a class of neural nets. This implies when these models are applied to the estimation of $(g_{10}, g_{00}, e_{10}, p_0)^\top$, Assumptions C.5.A and C.5.B hold, so our transported GATE estimator is asymptotically normal and Assumption 2.4 is satisfied.

Constructing Confidence Intervals for the Extrapolated Effects In our experimental setup, the data generating distribution for all observational studies is identical, so no transportation of effects is required, which enables the application of existing results. We use the doubly-robust score [46, 45] as the signal for the conditional average treatment effect,

$$Y(\eta) = \mu(1, X) - \mu(0, X) + \frac{A(Y - \mu(1, X))}{s(X)} - \frac{(1 - A)(Y - \mu(0, X))}{1 - s(X)}, \tag{18}$$

where $\eta := (\mu, s)$, and $\mu(A, X) := \mathbb{E}[Y|A, X]$, and $s(X) := \mathbb{P}(A = 1 \mid X)$. We use a multi-layer perceptron (MLP) regressor as a plug-in estimate $\hat{\mu}$ of μ, and ℓ_2-regularized logistic regression as a plug-in estimate $\hat{s}$ of s, with hyperparameters described in Section F.

Following example 2.2 from Semenova and Chernozhukov [55], we approximate the conditional treatment effect with a linear combination of subgroup dummy variables $G = (G_0, G_1, G_2, G_3)^\top$, so the combination weights correspond to the GATEs $\tau(k) = (\tau(k)_0, \tau(k)_1, \tau(k)_2, \tau(k)_3)$. This amounts to regressing the estimated signal $\hat{Y}_i(\hat{\eta})$ with G. As long as the propensity score is bounded above and below away from 0 and 1 (Assumption 4.10(a) of Semenova and Chernozhukov [55]), and the convergence rates of the response surface and propensity score estimates are sufficiently fast (Assumption 4.11), Corollary 4.1 and a set of mild technical conditions justify Theorem 3.1 in Semenova and Chernozhukov [55], which gives a result on pointwise asymptotic normality for the regression coeffcients $\hat{\tau}(k) = (\hat{\tau}(k)_0, \hat{\tau}(k)_1, \hat{\tau}(k)_2, \hat{\tau}(k)_3) \in \mathbb{R}^4$, so that for any unit vector $\gamma \in \mathbb{R}^4$ where $\|\gamma\| = 1$,

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}\gamma^\top(\hat{\tau}(k) - \tau(k))}{\sqrt{\gamma^\top \Omega \gamma}} < t \right) - \Phi(t) \right| = 0$$

where Ω can be consistently estimated with Equation 2.5 in Semenova and Chernozhukov [55]

$$\hat{\Omega} = \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1} \left(\frac{1}{N_k} \sum_j G_j G_j^\top (\hat{Y}_j(\hat{\eta}) - G_j^\top \hat{\tau}(k))^2 \right) \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1}$$

Setting γ as 1 in the $(i+1)$th element and 0 elsewhere thus yields

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))}{\sqrt{\Omega_{ii}}} < t \right) - \Phi(t) \right| = 0$$

We therefore estimate $\hat{\sigma}_i^2(k)$, the variance of $\hat{\tau}_i(k)$, with $\hat{\Omega}_{ii}$, and as this converges in probability to Ω_{ii}, the asymptotic normality of the above follows via Slutsky's theorem.

D Proofs

D.1 Proofs for propositions and theorems

Proposition 2.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 2.2 and 2.4 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1 - \rho)N$ for $\rho \in (0, 1)$. Define the test statistic*

$$\hat{T}_N(k, i) := \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{3}$$

where $\hat{s}^2 := \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) := \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k, i) \xrightarrow{d} \mathcal{N}(0, 1)$.

Proof. As $N \to \infty$, we have it that

$$\sqrt{\rho N}(\hat{\tau}_i(k) - \tau_i(k)) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(k))$$
$$\sqrt{(1-\rho)N}(\hat{\tau}_i(0) - \tau_i) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(0))$$

where we have written ρN in place of N_k, and similarly for N_0. By Slutsky's theorem, we can multiply by the constants $\rho^{-1/2}$ and $(1-\rho)^{-1/2}$ to get both results in terms of $\sqrt{N}$. We can then use independence of $\hat{\tau}(k), \hat{\tau}(0)$ to write that

$$\sqrt{N} \underbrace{\begin{pmatrix} \hat{\tau}_i(k) - \tau_i(k) \\ \hat{\tau}_i(0) - \tau_i \end{pmatrix}}_{Z-\theta} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_i^2(k)/\rho & 0 \\ 0 & \sigma_i^2(0)/(1-\rho) \end{bmatrix} \right).$$

We now apply the Delta method. Let $Z = (\hat{\tau}_i(k), \hat{\tau}_i(0))$ denote the (column) vector of estimates, and similarly let $\theta = (\tau_i(k), \tau_i)$. Letting $f(X) = X_1 - X_2$, we can argue that

$$\sqrt{N}(Z - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma) \implies \sqrt{N}(f(Z) - f(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \nabla f(\theta)^\top \Sigma \, \nabla f(\theta)\right),$$

where the resulting variance is given by

$$\nabla f(\theta)^\top \Sigma \, \nabla f(\theta) = \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho},$$

and $f(Z) - f(\theta) = \tau_i(k) - \tau_i - \mu_i(k)$.

$$\sqrt{N}(\tau_i(k) - \tau_i - \mu_i(k)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho}\right),$$

and accordingly that

$$\frac{\tau(k)_i - \tau_i - \mu_i(k)}{\sqrt{\frac{\sigma_i^2(k)}{N_k} + \frac{\sigma_i^2(0)}{N_0}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where this also holds (by Slutsky's theorem) with $\sigma^2(k)_i$ and $\sigma_i^2(0)$ replaced by their empirical estimates, which converge in probability. □

Theorem 3.1 (Properties of Algorithm 1)**.** *Under Assumptions 2.1 and 2.2, the output of Algorithm 1 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. *Under Assumptions 2.3 and 2.4, for each $i \in \mathcal{I}_O$,*

$$\lim_{N \to \infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{6}$$

2. *Under Assumption 2.4, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,*

$$\lim_{N \to \infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{7}$$

Proof. **(1)** By asymptotic normality and consistency of each dimension of $\tau(k)$, the test statistic $\hat{T}_N(k, i)$ converges in distribution to $\mathcal{N}(0, 1)$. As a result, for each $i \in \mathcal{I}_R$, the probability that $\left|\hat{T}_N(k, i)\right| > z_{\alpha/(4|\mathcal{I}_R|)}$ converges to $\alpha/(2\,|\mathcal{I}_R|)$. By an application of the union bound, the probability that this occurs for any $i \in \mathcal{I}_R$ is bounded by $\alpha/2$. Similarly, by the assumed properties of $\tau(k)$, the probability that the confidence interval $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ fails to capture the true value of τ_i converges to $\alpha/2$. By another application of the union bound, for each $i \in \mathcal{I}_O$, the probability that either $\tau(k)$ is not selected or τ_i is not contained in the interval is upper bounded by α. The result follows.

(2) By asymptotic normality of each $\tau(k)$, the power calculation in Equation (4) holds, and as $N \to \infty$, the probability of rejecting the null hypothesis converges to zero as $\sigma_{k,0}^2$ becomes arbitrarily large, which occurs as both $N_k, N_0 \to \infty$. □

Proposition C.1. *In the setting of Section B, under Assumptions B.1 and B.2, the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S=0, G=i] = \frac{\mathbb{P}(S=0)}{\mathbb{P}(S=0 \mid G=i)} \mathbb{E}[Y^a(\eta_0, \pi_0) \mid G=i], \tag{14}$$

where $Y^a(\eta, \pi)$ is defined as in Equation 15.

$$Y^a(\eta, \pi) := \frac{1}{\pi}\left(\mathbf{1}\{S=1, A=a\} \cdot \frac{1-p(X)}{p(X)e_a(X)} \cdot \{Y - g_a(X)\} + (1-S)g_a(X)\right) \tag{15}$$

where $\eta := (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi := \mathbb{P}(S=0)$ with true value π_0, $g_a(X) := \mathbb{E}[Y \mid A=a, S=1, X]$, $p(X) := \mathbb{P}(S=1 \mid X)$, and $e_a(X) := \mathbb{P}(A=a \mid S=1, X]$.

Proof. First, we can observe by standard arguments that the conditional expectation of $Y^a(\eta_0, \pi_0)$ given X is given by the following

$$\mathbb{E}[Y^a(\eta_0, \pi_0) \mid X=x] = \mathbb{E}\left[\frac{1-S}{\mathbb{P}(S=0)} g_{a0}(X) \middle| X=x\right],$$

because the first term in Equation (15) is mean-zero conditioned on $X=x$. This follows by the law of total expectation: for any event where $S=1, A=a$ does not hold, the first term is zero due to the indicator, and for any other event $S=1, A=a, X=x$, the first term is mean-zero, since the first term becomes a constant (determined by $S=1, A=a, X=x$) times a mean-zero random variable $Y - \mathbb{E}[Y \mid A=a, S=1, X=x]$.

As a result, we can write that

$$\begin{aligned}
&\mathbb{E}[Y^a(\eta_0, \pi_0) \mid G=i] \\
&= \mathbb{E}\left[\frac{1-S}{\mathbb{P}(S=0)} \mathbb{E}[Y \mid A=a, S=1, X] \middle| G=i\right] \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \sum_s \mathbf{1}\{s=0\} \mathbb{E}[Y \mid A=a, S=1, x] p(s, x \mid G=i) dx \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \sum_s \mathbf{1}\{s=0\} \mathbb{E}[Y_a \mid S=1, x] p(s, x \mid G=i) dx && \text{Assumption B.1}, Y_a \perp\!\!\!\perp A \mid X, S=1 \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \sum_s \mathbf{1}\{s=0\} \mathbb{E}[Y_a \mid S=0, x] p(s, x \mid G=i) dx && \text{Assumption B.2}, Y_a \perp\!\!\!\perp S \mid X \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(S=0, x \mid G=i) dx \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) \mathbb{P}(S=0 \mid G=i) dx \\
&= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) dx \\
&= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0, G=i] p(x \mid S=0, G=i) dx && X=x \Rightarrow G=i, \forall x : p(x \mid G=i) > 0 \\
&= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \mathbb{E}[Y_a \mid S=0, G=i]
\end{aligned}$$

and the result follows from dividing both sides by the first term on the right-hand side, which we can observe is equivalent to multiplying both sides by

$$\frac{\mathbb{P}(S=0)}{\mathbb{P}(S=0 \mid G=i)} = \frac{\mathbb{P}(S=0)\mathbb{P}(G=i)}{\mathbb{P}(S=0, G=i)} = \frac{\mathbb{P}(G=i)}{\mathbb{P}(G=i \mid S=0)} \tag{19}$$

□

D.2 Asymptotic normality of cross-fitted transported GATE estimators

Theorem C.1. *Suppose Assumptions C.1 to C.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$, where $\tilde{Y}$ is defined in Equation 16 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Proof sketch: Our strategy for the proof consists of two stages. First, we show that if the nuisance function is known to be η_0 and plugged into the estimator as $\mathbb{E}[\tilde{Y}(\eta_0, \hat{\pi}_g)|G = i]$, the resulting estimator, which we later refer to as the oracle estimator, is asymptotically normal. Second, we show that even if the true nuisance function is not known, as long as we have an estimator, $\hat{\eta}$, of the nuisance function that follows certain properties, the resulting estimator $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$ converges to the oracle estimator in probability. Then, by Slutsky's Theorem, the resulting estimator is also asymptotically normal.

Before diving into the first stage of the proof, we introduce additional notation to reflect the cross-fitting nature of our GATE estimator. Let the combined sample size of the observational study and RCT be N with sample indices $[N] := \{1, 2, ..., N\}$. We denote $(I_m)_{m=1}^M$ as a M-fold random partition of $[N]$, so that each fold has size $N_M = N/M$. The plug-in nuisance function estimate for the mth fold, $\hat{\eta}_m$, is then estimated from the rest of the folds $I_m^c := [N] \backslash I_m$. For brevity, we denote the size of the rest of the folds as $N_M^c = N - N/M$.

We now restate the definition of the treatment effect signal $\tilde{Y}_j(\eta, \pi_g) = \tilde{Y}_j((g_1, g_0, e_1, p)^\top, \pi_g)$:

$$\tilde{Y}_j(\eta, \pi_g) := \tilde{Y}_j^1(\eta, \pi_g) - \tilde{Y}_j^0(\eta, \pi_g)$$
$$\tilde{Y}_j^a(\eta, \pi_g) := \frac{1}{\pi_g(G_i)} \left(\mathbf{1}\{S_j = 1, A_j = a\} \cdot \frac{1 - p(X_j)}{p(X_j)e_a(X_j)} \cdot \{Y_j - g_a(X_j)\} + (1 - S_j)g_a(X_j) \right)$$

In the remainder of the development, we will drop the subscript j, which represents one of the N samples, for conciseness.

Stage 1 — *Proving the asymptotic normality of the oracle estimator*

For brevity, we define the following unweighted signal:

Definition D.1 (Unweighted signal functional)**.**

$$\begin{aligned}\mathcal{Y}(\eta) &= \pi_g(G)\tilde{Y}(\eta, \pi_g) \\ &= \pi_g(G)(\tilde{Y}^1(\eta, \pi_g) - \tilde{Y}^0(\eta, \pi_g)) \\ &= (1 - S)(g_1(X) - g_0(X)) + S\frac{1 - p(X)}{p(X)} \frac{(A - e_1(X))(Y - g_A(X))}{e_1(X)e_0(X)}\end{aligned}$$

From the proof of Proposition C.1, we have the following identities for the unweighted signals:

Lemma D.1 (Conditional mean of unweighted (oracle) signal)**.** *The conditional mean of the unweighted (oracle) signal is equivalent to the following:*

$$\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] = \tau_i \pi_g(i)$$
$$\mathbb{E}[\mathcal{Y}(\eta_0)|G = i, S = 0] = \tau_i$$

.

Proof. First, we have,

$$\begin{aligned}\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] &= \mathbb{E}[\pi_g(G)\tilde{Y}(\eta_0, \pi_g)|G = i] \\ &= \mathbb{E}[\pi_g(i)\tilde{Y}(\eta_0, \pi_g)|G = i] \\ &= \pi_g(i)\mathbb{E}[\tilde{Y}(\eta_0, \pi_g)|G = i] \\ &= \tau_i \pi_g(i)\end{aligned}$$

Next, using Definition D.1 of the unweighted signal functional and the fact that we condition on $S = 0$, we have,

$$\mathbb{E}[\mathcal{Y}(\eta_0)|G = i, S = 0] = \mathbb{E}\left[g_{10}(X) - g_{00}(X)|G = i, S = 0\right],$$

which is τ_i as desired. □

In addition, we can rewrite our estimator $\hat{\tau}_i := \mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$ with the unweighted signals:

$$\begin{aligned}
\hat{\tau}_i = \mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i] &= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\tilde{Y}(\hat{\eta}_m, \hat{\pi}_g)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\frac{1}{\hat{\pi}_g(G_j)}\mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)}, \quad \text{from } \textbf{Def. D.1} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\frac{1}{\hat{\pi}_g(i)}\mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{1}{\hat{\pi}_g(i)} \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\mathcal{Y}(\hat{\eta}_m)}{\frac{\sum_j \mathbf{1}(G_j = 1, S_j = 0)}{\sum_j \mathbf{1}(G_j = i)} \sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)\mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = 1, S_j = 0)}
\end{aligned}$$

Now, using the above expression, we can define the oracle estimator, where we *know* the true value of η, which is η_0:

Definition D.2 (Oracle GATE Estimator)**.**

$$\hat{\tau}_{i0} := \frac{\sum_j \mathbf{1}(G_j = i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j = i, S_j = 0)}$$

To show the asymptotic distribution of the oracle GATE estimator, we restate several assumptions:

Assumption C.1 (Observational dataset covers the whole support of covariates)**.**

$$\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0$$

Assumption C.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x)|G = i, S = 0] = \sigma^2_{\tau i} < \infty$$

Assumption C.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption C.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2|X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

These assumptions ensure that the oracle signals have finite conditional variance, which we prove in the following lemma.

Lemma D.2 (Finite conditional variance of unweighted oracle signal)**.** *Under Assumptions C.1 - C.4, we have that,*

$$var[\mathcal{Y}(\eta_0)|G = i] := \sigma_i^2 < \infty, \forall i \in [d]$$

Proof.

$$
\begin{aligned}
& var[\mathcal{Y}(\eta_0)|G=i] \\
=& \mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i] - [\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]]^2 \\
=& \mathbb{E}\Bigg[\Bigg((1-S)(g_{10}(X)-g_{00}(X)) + && \text{Lem. } D.1 \\
& \quad S\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\Bigg)^2\Bigg|G=i\Bigg] - \pi_g(i)^2\tau_i^2 \\
=& \Bigg\{\mathbb{E}\left[(1-S)(g_{10}(X)-g_{00}(X))^2\middle|G=i\right] + && \text{since } S\in\{0,1\} \\
& \quad \mathbb{E}\left[S\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i\right]\Bigg\} - \pi_g(i)^2\tau_i^2 \\
=& \Bigg\{\mathbb{E}\left[(g_{10}(X)-g_{00}(X))^2\middle|G=i,S=0\right]\pi_g(i) + \\
& \quad \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i,S=1\right](1-\pi_g(i))\Bigg\} - \pi_g(i)^2\tau_i^2 \\
=& \Bigg\{\left[var\left[g_{10}(X)-g_{00}(X)|G=i,S=0\right]+\tau_i^2\right]\pi_g(i) + \\
& \quad \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i,S=1\right](1-\pi_g(i))\Bigg\} - \pi_g(i)^2\tau_i^2 \\
<& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\mathbb{E}[(Y-g_{A0}(X))^2|G=i,S=1]}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 && \text{Asmp. } C.1-C.3 \\
\leq& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\bar{\sigma}^2}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 < \infty && \text{Asmp. } C.4
\end{aligned}
$$

□

Now, using the above lemmas, we are ready to prove the main result of stage 1 of the proof, stated below.

Proposition D.1 (Asymptotic normality of oracle GATE estimator)**.** *Under Assumptions C.1 - C.4,*

$$
\sqrt{N}(\hat{\tau}_{i0}-\tau_i) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2 - \pi_g(i)(1-\pi_g(i))\tau_i^2}{\pi_g(i)^2\mathbb{P}(G=i)}\right)
$$

Proof. We have that,

$$
\begin{aligned}
& \sqrt{N}(\hat{\tau}_{i0}-\tau_i) \\
=& \sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j=i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j=i,S_j=0)} - \tau_i\right) \\
=& \sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j=i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j=i,S_j=0)} - \frac{\sum_j \mathbf{1}(G_j=i,S_j=0)\tau_i}{\sum_j \mathbf{1}(G_j=i,S_j=0)}\right) \\
=& \sqrt{N}\frac{\sum_j \mathbf{1}(G_j=i)(\mathcal{Y}_j(\eta_0)-\tau_i\mathbf{1}(S_j=0))}{\sum_j \mathbf{1}(G_j=i,S_j=0)} \\
=& \frac{\sqrt{N}\frac{1}{N}\sum_j \mathbf{1}(G_j=i)(\mathcal{Y}_j(\eta_0)-\tau_i\mathbf{1}(S_j=0))}{\frac{1}{N}\sum_j \mathbf{1}(G_j=i,S_j=0)} \quad \begin{matrix} \xrightarrow{d} \mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i)) & \text{Proven below} \\ \xrightarrow{p} \mathbb{P}(G=i,S=0)=\mathbb{P}(G=i)\pi_g(i) & \text{From WLLN} \end{matrix} \\
\xrightarrow{d}& \mathcal{N}\left(0, \frac{\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2\mathbb{P}(G=i)}\right) && \text{Slutsky's lemma}
\end{aligned}
$$

In the above, we used the fact that,

$$\sqrt{N}\left[\frac{1}{N}\sum_j \mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))\right] \xrightarrow{d} \mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))$$

To show this, we observe that

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]\\
=&\mathbb{E}[\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0)|G=i]\mathbb{P}(G=i)\\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]-\tau_i\mathbb{E}[\mathbf{1}(S=0)|G=i])\mathbb{P}(G=i)\\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]-\tau_i\pi_g(i))\mathbb{P}(G=i)=0 && \text{Lem. } D.1
\end{aligned}$$

$$\begin{aligned}
&var[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]^2-(\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))])^2\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]^2\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))^2] && \mathbf{1}(G=i)\in\{0,1\}\\
=&\mathbb{E}[(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))^2|G=i]\mathbb{P}(G=i)\\
=&\left\{\mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i]+\tau_i^2\mathbb{E}[\mathbf{1}(S=0)|G=i]-2\tau_i\mathbb{E}[\mathcal{Y}(\eta_0)\mathbf{1}(S=0)|G=i]\right\}\mathbb{P}(G=i) && \mathbf{1}(S=0)\in\{0,1\}\\
=&\left\{var[\mathcal{Y}(\eta_0)|G=i]+(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i])^2+\tau_i^2\pi_g(i)\right.\\
&\qquad\left.-2\tau_i\pi_g(i)\mathbb{E}[\mathcal{Y}(\eta_0)|G=i,S=0]\right\}\mathbb{P}(G=i)\\
=&\left\{\sigma_i^2+\tau_i^2\pi_g(i)^2+\tau_i^2\pi_g(i)-2\tau_i^2\pi_g(i)\right\}\mathbb{P}(G=i) && \text{Asmp. } D.2, \text{Lem. } D.1\\
=&(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i)<\infty
\end{aligned}$$

Therefore, from central limit theorem,

$$\sqrt{N}\left[\frac{1}{N}\sum_j \mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))\right] \xrightarrow{d} \mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))$$

□

Stage 2 — *Proving the asymptotic normality of the cross-fitted estimator, $\tilde{Y}(\hat{\eta},\hat{\pi}_g)$*

With asymptotic normality of the oracle estimator shown above in Stage 1, we can show the asymptotic normality of the cross-fitted estimator (i.e. our estimator) by decomposing its error into the error of the oracle estimator and the difference between our estimator and the oracle estimator:

$$\begin{aligned}
\sqrt{N}(\hat{\tau}_i-\tau_i) &= \sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\sqrt{N}(\hat{\tau}_i-\hat{\tau}_{i0})\\
&= \sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\sqrt{N}\frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))}{\sum_j\mathbf{1}(G_j=i,S_j=0)}\\
&= \sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\frac{\sum_m\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))}{\frac{1}{N}\sum_j\mathbf{1}(G_j=i,S_j=0)}
\end{aligned}$$

The asymptotic distribution of the cross-fitted estimator therefore hinges on the asymptotic property of $\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))$, which in turn depends on the convergence property of the nuisance function estimate $\hat{\eta}_m$ and its influence on the signal $\mathcal{Y}$. We therefore restate the last required assumption governing the convergence properties of $\hat{\eta}_m$:

Assumption C.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers
- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)}\in\mathcal{T}_n)\geq 1-\epsilon_n$

- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions g_1, g_0, e_1, p, defined as follows:

$$\mathbf{g}_n \coloneqq \max_{a\in\{0,1\}} \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2}$$
$$\mathbf{e}_n \coloneqq \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2}$$
$$\mathbf{p}_n \coloneqq \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta\in\cup_{n=1}^{\infty}\mathcal{T}_n} \left(\max_{a\in\{0,1\}} \sup_{x\in\mathcal{X}} |g_a(x)| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{p(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_1(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_0(x)}\right| \right) = \bar{\mathcal{C}} < \infty$$

Based on the assumptions above, we have the following bounds on the convergence rate of the signals when the nuisance function estimates are in the high-probability neighborhood, $\mathcal{T}_n$:

Lemma D.3 (Bounds on bias of signal)**.** *Under Assumptions C.5.B and C.5.C*

$$\sqrt{n} \sup_{\eta\in\mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma D.4 (Bounds on MSE of signal)**.** *Under Assumptions C.1, C.3, C.4, C.5.A and C.5.C*

$$\sup_{\eta\in\mathcal{T}_n} \mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))\right|^2 = o_P(1)$$

which in turn implies that $\frac{1}{\sqrt{N}}\sum_{j\in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))$ converges to zero in probability:

Lemma D.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions C.1, C.3, C.4 and C.5,*

$$\frac{1}{\sqrt{N}} \sum_{j\in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0)) = o_P(1), \quad \forall m \in \{1, 2, ..., M\}$$

The proofs for Lemmas D.3 to D.5 are more labor-intensive and we defer these proofs to later subsections. Based on these lemmas, we arrive at the main result of Stage 2.

Theorem D.1 (Asymptotic normality of the cross-fitted transported GATE estimator)**.** *Under Assumptions C.1 - C.5,*

$$\sqrt{N}(\hat{\tau}_i - \tau_i) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2 \mathbb{P}(G=i)}\right)$$

Proof.

$$\begin{aligned}
&\sqrt{N}(\hat{\tau}_i - \tau_i) \\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}(\hat{\tau}_i - \hat{\tau}_{i0}) \\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}\frac{\sum_m \sum_{j\in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} \\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \frac{\sum_m \frac{1}{\sqrt{N}} \sum_{j\in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0)) \xrightarrow{p} 0}{\frac{1}{N}\sum_j \mathbf{1}(G_j = i, S_j = 0) \xrightarrow{p} \mathbb{P}(G = i, S = 0)} && \text{Lem. } D.5 \quad \text{WLLN} \\
\xrightarrow{d}&\mathcal{N}\left(0, \frac{\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2 \mathbb{P}(G=i)}\right) && \text{Prop. } D.1\text{, Slutsky's lemma}
\end{aligned}$$

□

Note that Theorem D.1 is simply Theorem C.1, which is the primary result of this section, with the variance explicitly stated. Thus, Theorem C.1 is proven.

D.3 Proof for Lemmas D.3 and D.4

First, we prove Lemmas D.3 and D.4, which will be necessary for Lemma D.5. Recall that Lemma D.5 was essential for the proof of the asymptotic normality result in Theorem D.1.

Lemma D.3 (Bounds on bias of signal)**.** *Under Assumptions C.5.B and C.5.C*

$$\sqrt{n} \sup_{\eta \in \mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma D.4 (Bounds on MSE of signal)**.** *Under Assumptions C.1, C.3, C.4, C.5.A and C.5.C*

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E} \left|\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))\right|^2 = o_P(1)$$

Proof. We first define partial unweighted signal functionals for the two counterfactual outcomes

Definition D.3 (Partial unweighted signal functionals)**.**

$$\begin{aligned}
\mathcal{Y}^1(\eta) &\coloneqq \left[(1-S)g_1(X) + S\frac{1-p(X)}{p(X)}\frac{A(Y-g_1(X))}{e_1(X)}\right] \\
\mathcal{Y}^0(\eta) &\coloneqq \left[(1-S)g_0(X) + S\frac{1-p(X)}{p(X)}\frac{(1-A)(Y-g_0(X))}{e_0(X)}\right] \\
\Rightarrow \mathcal{Y}(\eta) &= \mathcal{Y}^1(\eta) - \mathcal{Y}^0(\eta)
\end{aligned}$$

At a high level, we will prove the above lemmas by decomposing the errors of signal functionals into simpler terms that can be bounded by standard concentration inequalities. This idea will be repeated for both the bias and MSE of the signals. To simplify the analysis, we can split up the unweighted signal into "partial signals" (for the treatment and control groups). Therefore, we set out to show the following lemmas:

Lemma D.6 (Bounds on bias of partial signal)**.** *Under Assumptions C.5.B and C.5.C*

$$\begin{aligned}
\sqrt{n} \sup_{\eta \in \mathcal{T}_n} \left|\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))]\right| &= o_P(1) \\
\sqrt{n} \sup_{\eta \in \mathcal{T}_n} \left|\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}^0(\eta) - \mathcal{Y}^0(\eta_0))]\right| &= o_P(1)
\end{aligned}$$

Lemma D.7 (Bounds on MSE of partial signal)**.** *Under Assumptions C.1, C.3, C.4, C.5.A and C.5.C*

$$\begin{aligned}
\sup_{\eta \in \mathcal{T}_n} \mathbb{E} \left|\mathbf{1}(G = i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right|^2 &= o_P(1) \\
\sup_{\eta \in \mathcal{T}_n} \mathbb{E} \left|\mathbf{1}(G = i)(\mathcal{Y}^0(\eta) - \mathcal{Y}^0(\eta_0))\right|^2 &= o_P(1)
\end{aligned}$$

In the following subsections, we prove the $\mathcal{Y}_1$ part of Lemmas D.6 and D.7. The $\mathcal{Y}_0$ part will follow by symmetry. First, we further define $\eta(X) = \eta_0(X) + \delta_\eta(X)$, in detail:

$$\begin{aligned}
g_1(X) &= g_{10}(X) + \delta_{g_1}(X) \\
p(X) &= p_0(X) + \delta_p(X) \\
e(Z) &= e_0(X) + \delta_e(X)
\end{aligned}$$

so that (omitting the parameter X for brevity),

$$
\begin{aligned}
&\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0) \\
=&\left[(1-S)(g_{10}+\delta_{g_1}) + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e}\right] - \left[(1-S)g_{10} + \frac{1-p_0}{p_0}\frac{SA(Y-g_{10})}{e_0}\right] \\
=&(1-S)\delta_{g_1} + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e} - \frac{1-p_0}{p_0}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0} - \frac{1-p_0}{p_0}\frac{SA}{e_0}\delta_{g_1} \\
=&\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} + \left(\frac{1-p_0-\delta_p}{(p_0+\delta_p)(e_0+\delta_e)} - \frac{1-p_0}{p_0 e_0}\right)SA(Y-g_{10}-\delta_{g_1}) \\
=&\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} - \frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}SA(Y-g_{10}-\delta_{g_1}) \\
=&\underbrace{\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}}_{S_1} - \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}SA(Y-g_{10})}_{S_2} \\
&\qquad + \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}SA\delta_{g_1}}_{S_3} \\
:=&S_1 - S_2 + S_3
\end{aligned}
$$

D.3.1 Proof for Lemma D.6

For Lemma D.6 we want to bound

$$
\begin{aligned}
\left|\mathbb{E}\left[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right]\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0)\right)\middle|X\right]\right]\right| \\
&= \left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}\left[S_1 - S_2 + S_3|X\right]\right]\right| && G \text{ is a function of } X \\
&= \left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}\left[S_1|X\right] - \mathbb{E}\left[S_2|X\right] + \mathbb{E}\left[S_3|X\right]\right)\right]\right|
\end{aligned}
$$

For the term $\mathbb{E}[S_1|X]$,

$$
\begin{aligned}
&\mathbb{E}[S_1|X] \\
=&\mathbb{E}\left[\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\middle|X\right] \\
=&\left((1-\mathbb{E}[S|X]) - \frac{1-p_0}{p_0}\frac{\mathbb{E}[SA|X]}{e_0}\right)\delta_{g_1} && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\left((1-p_0) - \frac{1-p_0}{p_0}\frac{p_0 e_0}{e_0}\right)\delta_{g_1} = 0 && p_0(X) = \mathbb{P}[S=1|X] \\
& && e_0(X) = \mathbb{P}[A=1|S=1,X]
\end{aligned}
$$

For the term $\mathbb{E}[S_2|X]$

$$
\begin{aligned}
&\mathbb{E}[S_2|X] \\
=&\mathbb{E}\left[\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}SA(Y-g_{10})\middle|X\right] \\
=&\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}\mathbb{E}[SA(Y-g_{10}) \mid X] && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0 e_0}\cdot 0 = 0 && g_{10}(X) = \mathbb{E}[Y|S=1,A=1,X]
\end{aligned}
$$

For the term $\mathbb{E}[S_3|X]$,

$$
\begin{aligned}
&\mathbb{E}[S_3|X] \\
=&\mathbb{E}\left[\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0} SA\delta_{g_1}\middle| X\right] \\
=&\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\mathbb{E}\left[SA|X\right]\delta_{g_1} && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0} p_0e_0\delta_{g_1} && p_0(X) = \mathbb{P}[S=1|X] \\
& && e_0(X) = \mathbb{P}[A=1|S=1,X] \\
=&\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\mathbf{1}(G=i)[\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0)]\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}[S_1|Z] - \mathbb{E}[S_2|Z] + \mathbb{E}[S_3|Z]\right)\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right]\right|^2 \\
\leq&\left(\mathbb{E}\left|\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right|\right)^2 && |EA| \leq E|A| \\
\leq&\left(\mathbb{E}\left|\mathbb{E}[S_3|Z]\right|\right)^2 \\
=&\left(\mathbb{E}\left|\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 \\
\leq&\left(\mathbb{E}\left|\frac{e_0\delta_p}{pe}\delta_{g_1}\right| + \mathbb{E}\left|\frac{(1-p_0)p_0\delta_e}{pe}\delta_{g_1}\right| + \mathbb{E}\left|\frac{(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 && \text{Triangular ineq.} \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}\left|\delta_p\delta_{g_1}\right| + \mathbb{E}\left|\delta_e\delta_{g_1}\right| + \mathbb{E}\left|\delta_p\delta_e\delta_{g_1}\right|\right)^2 && \text{Assmp. } C.5.C \\
=&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}| + \mathbb{E}|\delta_e\delta_{g_1}| + \frac{1}{2}\mathbb{E}|\delta_e||\delta_p\delta_{g_1}| + \frac{1}{2}\mathbb{E}|\delta_p||\delta_e\delta_{g_1}|\right)^2 \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}| + \mathbb{E}|\delta_e\delta_{g_1}| + \frac{1}{2}\mathbb{E}|\delta_p\delta_{g_1}| + \frac{1}{2}\mathbb{E}|\delta_e\delta_{g_1}|\right)^2 && |\delta_p|, |\delta_e| \leq 1 \\
=&\frac{9}{4}\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}| + \mathbb{E}|\delta_e\delta_{g_1}|\right)^2 \\
\leq&\frac{9}{4}\bar{\mathcal{C}}^4\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2} + \sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right)^2 && \text{Hölder's ineq.}
\end{aligned}
$$

So we have,

$$
\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}\mathbf{1}(G=i)[Y^1(\eta) - Y^1(\eta_0)]\right| \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\frac{3}{2}\bar{\mathcal{C}}^2\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2} + \sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right) \\
\leq&\frac{3}{2}\bar{\mathcal{C}}^2\sqrt{n}\mathbf{g}_N\left(\mathbf{p}_N + \mathbf{e}_N\right) && \text{Assump. } C.5 \\
=&o_P(1) && \text{Assump. } C.5.B
\end{aligned}
$$

D.3.2 Proof for Lemma D.7

Here we first place bounds on $\mathbb{E}S_1^2, \mathbb{E}S_2^2$ and $\mathbb{E}S_3^2$ for future use. For the term $\mathbb{E}S_1^2$, we have,

$$
\begin{aligned}
&\mathbb{E}S_1^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\right)^2\middle| X\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&=\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\left(\frac{1-S}{1-p_0}-\frac{SA}{p_0e_0}\right)^2\middle|X\right]\delta_{g_1}^2\right] && \eta_0,\delta_\eta \text{ are functions of } X\\
&=\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{(1-S)^2}{(1-p_0)^2}+\frac{(SA)^2}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && S\in\{0,1\}\\
&=\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{1-S}{(1-p_0)^2}+\frac{SA}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && 1-S, SA\in\{0,1\}\\
&=\mathbb{E}\left[(1-p_0)^2\left(\frac{1}{1-p_0}+\frac{1}{p_0e_0}\right)\delta_{g_1}^2\right] && \begin{aligned}p_0(X)&=\mathbb{P}[S=1|X]\\ e_0(X)&=\mathbb{P}[A=1|S=1,X]\end{aligned}\\
&\le\frac{2}{\varepsilon_p\varepsilon_e}\mathbb{E}\delta_{g_1}^2 && \text{Assmp. } C.1, C.3
\end{aligned}
$$

We can similarly bound $\mathbb{E}S_2^2$,

$$
\begin{aligned}
&\mathbb{E}S_2^2\\
&=\mathbb{E}\left[\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})\right)^2\middle|X\right]\right]\\
&=\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\mathbb{E}\left[(SA(Y-g_{10}))^2\middle|X\right]\right] && \eta_0,\delta_\eta \text{ are functions of } X\\
&=\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\right. && S,A\in\{0,1\}\\
&\qquad\qquad\left.\mathbb{E}\left[(Y-g_{10})^2\middle|X,S=1,A=1\right]\mathbb{P}(S=1,A=1|X)\right]\\
&\le E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\bar{\sigma}^2\mathbb{P}(S=1,A=1|X)\right] && \text{Assmp. } C.4\\
&=E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pep_0e_0}\right)^2\bar{\sigma}^2p_0e_0\right] && \begin{aligned}p_0(X)&=\mathbb{P}[S=1|X]\\ e_0(X)&=\mathbb{P}[A=1|S=1,X]\end{aligned}\\
&\le\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left|e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e\right|^2 && \text{Assmp. } C.1, C.3, C.5.C\\
&\le\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[e_0|\delta_p|+(1-p_0)p_0|\delta_e|+(1-p_0)|\delta_p\delta_e|\right]^2 && \text{Triangular ineq.}\\
&\le\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+|\delta_p\delta_e|\right]^2\\
&=\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+\frac{1}{2}|\delta_p||\delta_e|+\frac{1}{2}|\delta_p||\delta_e|\right]^2\\
&\le\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+\frac{1}{2}|\delta_p|+\frac{1}{2}|\delta_e|\right]^2 && |\delta_p|,|\delta_e|\le 1\\
&=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|\right]^2\\
&=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left[\mathbb{E}\delta_p^2+\mathbb{E}\delta_e^2+2\mathbb{E}|\delta_p||\delta_e|\right]\\
&\le\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left[\mathbb{E}\delta_p^2+\mathbb{E}\delta_e^2+2\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_e^2}\right] && \text{Cauchy-Schwartz}\\
&=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left(\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right)^2
\end{aligned}
$$

Finally, we bound $\mathbb{E}S_3^2$,

$$
\begin{aligned}
&\mathbb{E}S_3^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}\right)^2\middle|X\right]\right] \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\delta_{g_1}\right)^2\mathbb{E}\left[(SA)^2\middle|X\right]\right] && \eta_0,\delta_\eta \text{ are functions of } X \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\delta_{g_1}\right)^2\mathbb{E}\left[SA|X\right]\right] && SA\in\{0,1\} \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pep_0e_0}\delta_{g_1}\right)^2p_0e_0\right] && \begin{aligned}p_0(X)&=\mathbb{P}[S=1|X]\\ e_0(X)&=\mathbb{P}[A=1|S=1,X]\end{aligned} \\
\leq&\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left|(e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e)\,\delta_{g_1}\right|^2 && \text{Assmp. } C.1, C.3, C.5.C \\
=&\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e|^2|\delta_{g_1}|^2\right] \\
\leq&\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[(|e_0\delta_p|+|(1-p_0)p_0\delta_e|+|(1-p_0)\delta_p\delta_e|)^2\,|\delta_{g_1}|^2\right] && \text{Triangular ineq.} \\
\leq&\frac{9\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\delta_{g_1}^2 && 0\leq p_0,e_0,|\delta_p|,|\delta_e|\leq 1
\end{aligned}
$$

From the above, we have

$$
\begin{aligned}
&\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2 \\
=&\mathbb{E}|\mathbf{1}(G=i)||S_1-S_2+S_3|^2 \\
\leq&\mathbb{E}|S_1-S_2+S_3|^2 \\
\leq&\mathbb{E}(|S_1|+|S_2|+|S_3|)^2 && \text{Triangular ineq.} \\
=&\left[\mathbb{E}S_1^2+\mathbb{E}S_2^2+\mathbb{E}S_3^2+2\mathbb{E}|S_1S_2|+2\mathbb{E}|S_1S_3|+2\mathbb{E}|S_2S_3|\right] \\
\leq&\Big[\mathbb{E}S_1^2+\mathbb{E}S_2^2+\mathbb{E}S_3^2+ \\
&\quad 2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_2|^2}+2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_3|^2}+2\sqrt{\mathbb{E}|S_2|^2}\sqrt{\mathbb{E}|S_3|^2}\Big] && \text{Cauchy-Schwartz} \\
=&\left[\sqrt{\mathbb{E}S_1^2}+\sqrt{\mathbb{E}S_1^2}+\sqrt{\mathbb{E}S_1^2}\right]^2 \\
\leq&\left[\sqrt{\frac{2}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}}\left(\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right)+\sqrt{\frac{9\bar{C}^4}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2}\right]^2 \\
\leq&C\left[\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right]^2 && C:=\frac{\left(3\bar{C}^2+\sqrt{2}\right)^2}{\varepsilon_p\varepsilon_e}\vee\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}
\end{aligned}
$$

So we have,

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2 \\
\leq&C\sup_{\eta\in\mathcal{T}_n}\left[\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right]^2 \\
\leq&C\left[\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_p^2}+\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_e^2}\right]^2 \\
\leq&C\left[\mathbf{g}_n+\mathbf{p}_n+\mathbf{e}_n\right]^2 && \text{Assmp. } C.5 \\
\leq&9C\left[\mathbf{g}_n\vee\mathbf{p}_n\vee\mathbf{e}_n\right]^2=o_P(1) && \text{Assmp. } C.5.A
\end{aligned}
$$

D.3.3 Assembling the proofs for Lemmas D.3 and D.4

For Lemma D.3:

$$
\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))]| \\
=&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right| \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left\{\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right|+\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right|\right\} && \text{Triangular ineq.} \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right|+\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right|=o_P(1) && \text{Lem. } D.6
\end{aligned}
$$

For Lemma D.4:

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))\right|^2 \\
=&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))-\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2 \\
\leq&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left\{\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|+\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\right\}^2 && \text{Triangular ineq.} \\
=&\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ \\
&\qquad 2\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\Big\} \\
\leq&\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ && \text{Cauchy-Schwartz} \\
&\qquad 2\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2}\Big\} \\
\leq&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ \\
&\qquad 2\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2} \\
=&o_P(1) && \text{Lem. } D.7
\end{aligned}
$$

□

D.4 Proof for Lemma D.5

Now that we have shown Lemmas D.3 and D.4, it remains to show Lemma D.5.

Lemma D.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions C.1, C.3, C.4 and C.5,*

$$
\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))=o_P(1),\quad \forall m\in\{1,2,...,M\}
$$

Proof. First we observe

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0)) \\
=&\frac{1}{\sqrt{N}}\left[\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]+ \\
&\qquad\qquad\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))] \\
=&\underbrace{\frac{N_M}{\sqrt{N}}\left[\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right)-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]}_{R_1(m)}+
\end{aligned}
$$

$$\underbrace{\frac{N_M}{\sqrt{N}}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]}_{R_2(m)}$$

$$:= R_1(m) + R_2(m)$$

We define the event $\mathcal{E}_N$ as $\cap_k \left(\hat{\eta}_m \in \mathcal{T}_{N_M^c}\right)$, i.e. all M nuisance function estimates falling into the high-probability neighborhood where Lemmas D.3 and D.4 apply. From union bound,

$$1 - \mathbb{P}(\mathcal{E}_N) \leq \sum_k \mathbb{P}(\hat{\eta}_m \notin \mathcal{T}_{N_M^c}) \leq K\epsilon_{N_M^c} = o_P(1) \qquad \because \epsilon_n = o_P(1)$$

Conditional on $\mathcal{E}_N$ and the data complementary to fold m, which we denote as D_m, we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(|R_1(m)| \geq \epsilon | \mathcal{E}_N, D_m) \\
=&\mathbb{P}\left(\left|\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right) - \right.\right. \\
&\qquad\qquad \left.\left.\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right| \geq \frac{\sqrt{N}}{N_M}\epsilon \,\middle|\, \mathcal{E}_N, D_m\right) \\
\leq&\frac{N_M^2}{N\epsilon^2} var\left[\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0)) \,\middle|\, \mathcal{E}_N, D_m\right] && \text{Chebyshev ineq.} \\
=&\frac{N_M}{N\epsilon^2} var\left[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0)) | \mathcal{E}_N, D_m\right] \\
\leq&\frac{1}{M\epsilon^2}\mathbb{E}\left[(\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0)))^2 \,\middle|\, \mathcal{E}_N, D_m\right] \\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left[(\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0)))^2 \,\middle|\, D_m\right] && \hat{\eta}_m \in \mathcal{T}_{N_M^c} \text{ under } \mathcal{E}_N \\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left(\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))\right)^2 && \text{Fold } m \text{ is independent of } D_m \\
=&\frac{1}{M\epsilon^2}o_P(1) = o_P(1) && \text{Lem. } D.4
\end{aligned}$$

Also, conditional on $\mathcal{E}_N$ and D_m

$$\begin{aligned}
|R_2(m)| =& \left|\frac{N_M}{\sqrt{N}\sqrt{N_M^c}}\sqrt{N_M^c}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))] \,\middle|\, \mathcal{E}_N, D_m\right| \\
=&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\,|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))] | \mathcal{E}_N, D_m| \\
\leq&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\sup_{\eta\in\mathcal{T}_{N_M^c}}|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0)) | D_m]| && \hat{\eta}_m \in \mathcal{T}_{N_M^c} \text{ under } \mathcal{E}_N \\
=&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\sup_{\eta\in\mathcal{T}_{N_M^c}}|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))]| && \text{Fold } m \text{ is independent of } D_m \\
=&\frac{1}{\sqrt{M-1}}o_P(1) = o_P(1) && \text{Lem. } D.3
\end{aligned}$$

Which implies that, conditional on $\mathcal{E}_N$ and D_m, $R_1(m) + R_2(m) = o_P(1)$. So for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right| > \epsilon\right)$$

$$
\begin{aligned}
&=\mathbb{P}(|R_1(m)+R_2(m)|\geq \epsilon)\\
&=\mathbb{P}(|R_1(m)+R_2(m)|\geq \epsilon|\mathcal{E}_N)\mathbb{P}(\mathcal{E}_N)+\mathbb{P}(|R_1(m)+R_2(m)|\geq \epsilon|\bar{\mathcal{E}}_N)(1-\mathbb{P}(\mathcal{E}_N))\\
&\leq\mathbb{P}(|R_1(m)+R_2(m)|\geq \epsilon|\mathcal{E}_N)+(1-\mathbb{P}(\mathcal{E}_N))\\
&=\int \mathbb{P}(|R_1(m)+R_2(m)|\geq \epsilon|\mathcal{E}_N, D_m)d\mathbb{P}(D_m|\mathcal{E}_N)+(1-\mathbb{P}(\mathcal{E}_N))\\
&=o_P(1)+o_P(1)=o_P(1)
\end{aligned}
$$

and the lemma is proven. □

E Details on WHI Experiments

We assess our algorithm on clinical trial data and observational data available from the Women's Health Initiative (WHI). The RCTs were run by the WHI via 40 US clinical centers from 1993-2005 (1993-1998: enrollment + randomization; 2005: end of follow-up) on postmenopausal women aged 50-79 years, and the observational dataset was designed and run in parallel on a similar population. Note that this data is publicly available to researchers and requires only an application on BIOLINCC (https://biolincc.nhlbi.nih.gov/studies/whi_ctos/).

E.1 Data

WHI RCT – There are three clinical trials associated with the WHI. The RCT that we will be leveraging in this set of experiments is the Postmenopausal Hormone Therapy (PHT) trial, which was run on postmenopausal women aged 50-79 years who had an intact uterus. This trial included a total of $N_{HT} = 16608$ patients. The intervention of interest was a hormone combination therapy of estrogen and progesterone. Specifically, post-randomization, the treatment group was given 2.5 mg of medroxyprogesterone as well as 0.625 mg of estrogen a day. The control group was given a placebo. Finally, there are several outcomes that were tracked and studied in the principal analysis done on this trial [51]. These outcomes are of three broad categories: a) cardiovascular events, including coronary heart disease, which served as a primary endpoint b) cancer (e.g. endometrial, breast, colorectal, etc.), and c) fractures.

WHI OS – The observational study component of the WHI tracked the medical events and health habits of $N = 93676$ women. Recruitment for the study began in 1994 and participants were followed until 2005, i.e. a similar follow-up to the RCT. Follow-up was done in a similar fashion as in the RCT (i.e. patients would have annual visits, in addition to a "screening" visit, where they would be given survey forms to fill out to track any events/outcomes). Thus, the same outcomes, including cancers, fractures, and cardiovascular events, are tracked in the observational study.

E.2 Outcome

The outcome of interest in our analysis is a "global index", which is a summary statistic of several outcomes, including coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. Events or outcomes are tracked for each patient, and are recorded as "day of event/outcome" in the data, where the initial time-point for follow-up is the same for both the RCT and OS. At a high level, the "global index" is essentially the minimum "event day" when considering all the previously mentioned events.

We binarize the "global index," by choosing a time point, t, before the end of follow-up and letting $Y = 1$ if the observed event day is before t and $Y = 0$ otherwise. Thus, we are looking at whether the patient will experience the event within some particular period of time or not. We set $t = 7$ years. Note that we sidestep censorship of a patient before the threshold by defining the outcomes in the following way: $Y = 1$ indicates that a patient is observed to have the event before the threshold, and $Y = 0$ indicates that a patient is not observed to have the event before the threshold. We apply this binarization in the same way for both the RCT and OS. Extending our method to a survival analysis framing is beyond the scope of this paper, but an interesting direction for future work.

E.3 Intervention

Recall from above that the intervention studied in the RCT was 2.5mg of medroxyprogesterone + 0.625 mg of estrogen and the control was a placebo pill. The RCT was run as an "intention-to-treat" trial. To establish "treatment" and "control" groups in the OS, we leverage the annual survey data collected from patients and assign a patient to the treatment group if they confirm usage of both estrogen and progesterone in the first three years. A patient is assigned to the control group if they deny usage of both estrogen and progesterone in the first three years. We exclude a patient from the analysis if she confirms usage of one and not the other OR if the field in the survey is missing OR if they take some other hormone therapy. We end up with a total of $N_{obs} = 33511$ patients.

E.4 Data Processing + Covariates

We use only covariates that are measured both in the RCT and OS to simplify the analysis. Because this information is gathered via the same set of questionnaires, they each indicate the same type of covariate. In other words, there is consistency of meaning across the same covariates across the RCT and the OS. We end up with a total of 1576 covariates.

E.5 Details of Experimental Setup

We give a more detailed exposition of the steps in our experimental workflow, which were described in brief in the main paper.

- **Step 0**: *Replicate the principal results from the PHT trial, given in Table 2 of [51], using the WHI OS data.* In this step, we fit a doubly robust estimator of the style given in Appendix C.
- **Step 1**: *While treating the WHI OS dataset as the "unbiased" observational dataset (hence the need for Step 0), simulate additional "biased" observational datasets by inducing bias into the WHI OS.* We construct four additional "biased" datasets (for a total of five observational datasets, including the WHI OS dataset), where we use the following procedure to induce selection bias – of the people who were not exposed to the treatment and did not end up getting the event, we drop each person with some probability, p. We set $p = [0.1, 0.3, 0.5, 0.7]$ to get the four additional observational datasets.

  This type of selection bias may reflect the following clinical scenario: consider a patient who is relatively healthy who does not end up taking any hormone therapy. This patient might enroll initially in the OS, but may drop out or stop responding to the surveys. If the committee running the study does not explicitly account for this drop-out rate, then the resultant study will suffer from selection bias. [4] detail additional examples of selection bias that can occur in observational studies. Importantly, this part is the only part of our setup that involves any simulation. However, in order to properly evaluate our method, we need to know which datasets are biased and unbiased in our set. Thus, we opt to simulate the bias.
- **Step 2**: *Run our procedure over "multiple tasks," generating confidence intervals on the treatment effect for different subgroups.* To do so, we compile a list of covariates, taking both from [52] as well as covariates with high feature importance in both the propensity score model and response surface model from the estimator in Step 0. We generate all pairs from this list and use each pair to generate four subgroups. We treat two of the subgroups as validation subgroups and two of them as extrapolated subgroups in that we "hide" the RCT data in those subgroups when fitting our doubly robust transported estimator. (This gives us the benefit of knowing the RCT result for the extrapolated subgroups, which is useful in evaluation). Pairs that do not have enough support (threshold of 400 observations) in each group are removed. The total number of "tasks" (or covariate pairs) that we have is 592 (and therefore 2368 subgroups).
- **Step 3**: *Evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis for each of the covariate pairs.* Additionally, we evaluate an "oracle" method, which always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias) and reports the interval estimate computed on this study. To evaluate these methods, we will treat the RCT point estimates as "correct." For each, we compute the following metrics: **Length** – length of the confidence interval for the subgroup; **Coverage** – percentage of tasks

for which the method's interval covers the RCT point estimate; **Unbiased OS Percentage** – across all tasks, the percentage at which our approach retains the unbiased study after the falsification step.

Note that we utilize sample splitting when running the above procedure. Namely, we use 50% of the data as a "training" set, where we experiment with different classes of covariates and different types of bias, and then reserve 50% of the data as a "testing" set, on which we do the final run of the analysis and report results. All nuisance functions in the doubly robust estimator are fit with a Gradient Boosting Classifier with significant regularization. In practice, we found that any highly-regularized tree-based model works well.

E.6 Covariate List for Task Generation

Below is the list of covariates used to generate the tasks in **Step 2** of our experiment:

- ALCNOW (current alcohol user)
- BMI ≤ 30
- BLACK
- SMOKING (current smoker)
- DIAB (diabetes ever)
- HYPT (hypertension ever)
- BRSTFEED (breastfeeding)
- MSMINWK ≤ 106 (minutes of moderate to strenuous activity per week)
- BRSTBIOP (breast biopsy done)
- RETIRED
- EMPLOYED
- OC (oral contraceptive use ever)
- LIVPRT (live with husband or partner)
- MOMALIVE (natural mother still alive)
- LATREGION-Northern > 40 degrees north
- BKBONE (broke bone ever)
- NUMFALLS
- GRAVID (gravidity)
- AGE ≤ 64
- ANYMENSA ≤ 51 (age at last bleeding)
- MENOPSEA ≤ 50 (age at last regular period)
- MENO $\leq$51 (age at menopause)
- LSTPAPDY (days from randomization to last pap smear)
- BMI ≤ 27.7
- TMINWK ≤ 191 (minutes of recreational exercise per week)
- HEIGHT ≤ 161
- WEIGHT ≤ 72
- WAIST ≤ 86
- HIP ≤ 105
- WHR ≤ 0.81 (waist to hip ratio)
- TOTHCAT (HRT duration by category)
- MEDICARE (on medicare)
- HEMOGLBN ≤ 13

- PLATELET ≤ 244
- WBC ≤ 6
- HEMATOCR ≤ 40

F Details on Semi-Synthetic Experiment (Data Generation and Model Hyperparameters)

F.1 Data Generation

For each simulated dataset, we generate 1 RCT and K observational studies. The RCT is assumed to have covariate values identical to the IHDP dataset but is restricted to infants with married mothers. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = rn_0$. The covariate distribution of the observational studies are made different from the RCT by weighted sampling, with the relative weights set as

$$w = 0.8^{\mathbf{1}(\text{male infant})+\mathbf{1}(\text{mother smoked})+\mathbf{1}(\text{mother worked during pregnancy})}$$

Then, to introduce confounding (in the observational data), we generate m_c continuous confounders and m_b binary confounders. Each continuous confounder is drawn from a mixture of $0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(3,1)$ in the RCT and $(0.25+0.5A)\mathcal{N}(3,1)+(0.75-0.5A)\mathcal{N}(0,1)$ in the observational studies, where A is the treatment indicator. Similarly, each binary confounder is drawn from Bern(0.5) in the RCT and Bern$(0.25 + 0.5A)$ in the observational studies. In the following, we denote the covariate vector as $X \in \mathbb{R}^{m_x}$ where $m_x = 28$ is the number of covariates in the IHDP dataset, and the generated confounder vector as $Z \in \mathbb{R}^{(m_c+m_b)}$. For brevity, we also denote the vector $(A, X^\top)^\top$ as $\tilde{X}$.

For outcome simulation in the datasets, we modify *response surface B* from Hill [25] to account for additional confounding variables. We set the following counterfactual outcome distributions:

$$Y_0 \sim \mathcal{N}\left(\exp\left[\left(\tilde{X} + \frac{1}{2}\mathbf{1}\right)^\top \beta\right] + Z^\top\gamma, 1\right)$$
$$Y_1 \sim \mathcal{N}(\tilde{X}^\top\beta + Z^\top\gamma + \omega, 1),$$

where $\mathbf{1} \in \mathbb{R}^{(m_x+1)}$ is vector of ones, $\beta \in \mathbb{R}^{(m_x+1)}$ is a vector where each element is randomly sampled from $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, $\gamma \in \mathbb{R}^{(m_c+m_b)}$ is a vector where each element is randomly sampled from $(0.1, 0.2, 0.5, 0.75, 1)$ with uniform probability, and $\omega = 23$ is a constant chosen to limit the size of the GATEs. The observed outcome is then $Y := AY_1 + (1 - A)Y_0$. We then conceal a number of confounders, chosen in order from the highest to lowest weighted, from each observational study to mimic the scenario of unobserved confounding. The number of concealed confounders in each observational study is denoted as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$.

F.2 Hyperparameters

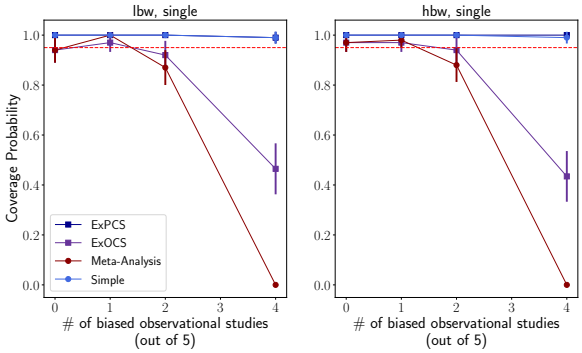
F.2.1 Logistic regression

Hyperparametes	Value set
Penalty type	ℓ_2
Penalty coefficient	$\{1, 0.1, 0.01, 0.001\}$

F.2.2 Multilayer perceptron regression

Hyperparametes	Value set
# of hidden layers and # of perceptrons	$[1, (100)], [2, (50, 50)], [2, (25, 25)]$
Activation function	ReLU, tanh
Solver	Adam
Alpha	$(1, 0.1, 0.01, 0.001, 0.0001)$
Learning rate	0.001
# of epochs	$(250, 500)$

G Additional Semi-Synthetic Experimental Results



Mean width of 95% confidence intervals (4 biased studies)				
	ExPCS	ExOCS	Meta	Simple
LS	5.29	3.55	2.34	5.51
HS	5.84	3.76	2.48	6.15

Figure 5: Coverage probabilities of confidence intervals shown as a function of the number of biased observational datasets (out of five). In the 4/5 biased studies case, the average interval widths for each approach is shown for two subgroups. We observe that *ExPCS* achieves the best balance of interval width and coverage.

Upsampling ratio	1.0	3.0	5.0	10.0
P(selecting biased study)	0.98	0.80	0.68	0.60

Table 2: P(selecting biased study) as a function of upsampling ratio

An analysis of including biased observational studies: In Figure 5, we study coverage probability and width of confidence intervals in the presence of biased studies. *Meta-Analysis* intervals approach zero coverage probability as the number of biased studies increases. Indeed, a fundamental assumption of this approach is that differences between estimates are only due to random variation, leading to poorer coverage probability when there are more biased studies. *ExOCS* allows for elimination of biased studies in principle through falsification, resulting in improved coverage. However, it does not maintain the desired threshold of coverage (95%), since biased estimators may still be included after falsification either due to chance or by being underpowered. Finally, *ExPCS* and *Simple Union* intervals have good coverage across the board, but as before, *ExPCS* results in narrower intervals.

Overall, we find that our method is robust to biased studies, yielding a good balance between coverage and width. In the case where one has adequate power, *ExOCS* could be a reasonable alternative to get narrower intervals for a sacrifice in coverage (even in the presence of biased studies). However, this implicitly assumes that an estimator consistent for the validation effects will be consistent for the extrapolated effects. If this assumption does not hold, then *ExOCS* will have poor coverage.

Biased estimator selection: In Table 2, we see that the probability of selecting the biased estimator goes down with increasing sample size of the observational studies, reflected by the increasing sample size ratio, r. This result validates our intuition that our method is more useful and results in more precise estimates of bias as we obtain more observational samples.

H Additional related work

Combining observational and experimental data Prior work has sought to combine RCTs and observational studies for the purpose of more precise estimation of treatment effects [49, 50], or for the purpose of generalizing or transporting estimates from RCTs to observational populations when overlap holds between the two (see Degtiar and Rose [16] for a recent review). In contrast, our work is motivated by settings where there are populations in the observational studies who are not at all represented in trials, e.g., due to a lack of eligibility. Kallus et al. [30] also seek to combine observational and experimental data to extrapolate beyond tfhe support of an RCT. They propose to learn a CATE function on observational data, and then learn a parametric additive correction term on the sample that overlaps between the RCT and observational data. In contrast to this approach, we

do not assume that confounding can be corrected for, and instead seek to choose an observational estimate (if one exists) that is already consistent for each sub-population.

Calibrating observational confidence intervals An alternative method for calibration of confidence intervals for observational studies makes use of negative controls [34], such as drug-outcome pairs known to have no causal relationship. Methods range using uses these negative controls to form an empirical null distribution for callibrated p-values[53], and Schuemie et al. [54] extend this approach to calibration of confidence intervals. These techniques have been used in several large-scale observational studies such as the LEGEND-HTN study for comparing antihypertensive drugs [59]. By contrast, our method does not assume the existence of negative controls, but instead uses a form of positive control (i.e., validation effects), one that is based on the same underlying treatment as the extrapolated effect.