Falsification before Extrapolation in Causal Effect Estimation

Anonymous Author(s)
Affiliation
Address
email

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 experiments based on the IHDP dataset, 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 [30, 12, 42]. In healthcare, when a target population has reasonable overlap with the inclusion criteria of RCTs, current clinical treatment guidelines rely primarily on RCTs [22, 21]. 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 subpopulation 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 [55]. The CoMMpass trial, an observational dataset, was also available in which the Carfilzomib regimen was given to both NDMM and RRMM patients [35]. 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 [32, 31].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 [23, 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,[1] 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 [41, 13, 14]. 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 [11, 52]) 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 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 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 $\{1, \dots, I\} \setminus \mathcal{I}_R$. 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 $i \in \{1, \dots, I\}$ and $j \in \{1, \dots, J\}$, i.e., all observational datasets ($D \geq 1$) have support for all groups.

[1] 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.

Definition 2.1 (Validation and Extrapolated Effects)**.** We define the group average treatment effect (GATE)[2] 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 [53, 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, where the remainder are observational estimators.[3] 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 a **estimator** of the resulting observational quantity which is asymptotically normal [40, 41, 39, 14, 15]. 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[4] 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 $\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))$.

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] We use this term in line with the literature [10, 27, 37, 52] and to distinguish it from the CATE function.

[3] 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.

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

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 [52].

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.[5] One approach, given in Chernozhukov et al. [11], is to use double machine learning (DML), which employ cross-fitting to produce estimates $\hat{\tau}(k)$ based on the doubly-robust score [44], 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 11, for additional details]. These results hold whether or not $\tau(k) = \tau$, as discussed in Footnote 9 of Chernozhukov et al. [11]. 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 [15] with the use of transported estimators [14] (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 [38] 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.[6] 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 [6, 28]

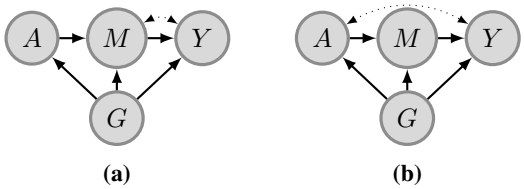


Figure 1: (Ex. 2.2) In (a), M and Y are confounded by unobservables (bi-directional dotted arrow). In (b), A and Y are confounded, but the causal effect is identified via front-door adjustment.

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 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).

[5] A rich literature focuses on establishing such results, beyond the approach in this example [2, 19, 57, 36, 3].

[6] 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.

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 exists estimators for the ATE in this setting, under mild additional assumptions [14, 13]. 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 [52], 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 59]. 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.

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 $\rho_k > 0$.*

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 $i \in \mathcal{I}_O$.

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 Experiments

4.1 Setup of Simulation

Performance assessement of our meta-algorithm on a real-world dataset can be challenging, since we typically lack access to ground-truth treatment effects. Therefore, we generate semi-synthetic RCTs and observational datasets, with covariates from a real-world experiment and simulated outcomes. We use the Infant Health and Development Program (IHDP), a randomized experiment on premature infants, assessing the effect of intensive high-quality child care and home visits from a trained provider on the infants' future cognitive performance [7]. Our data generation is based on the partial IHDP dataset used in [24], which includes $n_0 = 985$ observations with a binary treatment variable, along with 28 covariates. We construct a scenario where there are four subgroups of interest, 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. This mimics a scenario where only infants with married mothers are recruited into the RCT.

For each simulated dataset, we generate 1 RCT and K observational studies. The RCT has 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 $n = r \cdot n_0$. 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. We then introduce confounding in the observational data, generating m_c continuous confounders and m_b binary confounders. Finally, we simulate outcomes in each dataset,

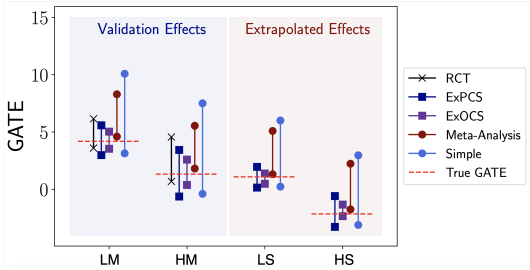


Figure 2: The confidence intervals for group average treatment effects (GATE) within the four subgroups output by our algorithm (*ExPCS*), our algorithm variant (*ExOCS*), random-effects meta-analysis on all observational studies (*Meta-Analysis*), simple union bound on all observational studies (*Simple*), 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

modifying the response surface given in Hill [24]. In our experiments, we may choose to conceal some number of 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 E. 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 E, and details of hyperparameter tuning can be found in Appendix C.

4.2 Implementation and Evaluation of Meta-Algorithm

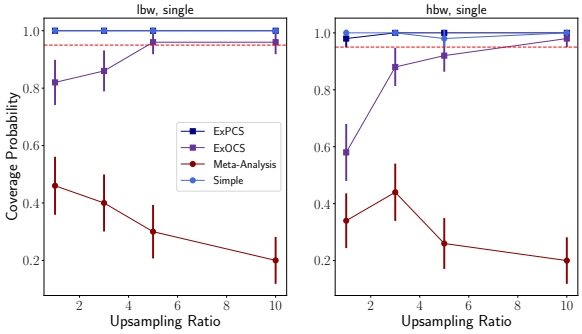
To implement Algorithm 1, we first obtain GATE estimates for the four subgroups and their estimated variances in each observational study, combining techniques from the DML and trasportability literature [52, 14]. Estimation details is shown in Appendices B and C. For the RCT, we stratify the data into the subgroups HM, LM and estimate the GATEs as the difference of mean outcome 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, and (2) the width of the confidence intervals. Ideally, the confidence intervals would have adequate coverage probability (at least $1 - \alpha$) and be as narrow as possible. 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 5, with heterogeneity variance estimated via the DerSimonian-Laird moment method [16]. This baseline is the current standard for aggregating observational study results. The second baseline, *Simple Union*, reports the union bounds of the confidence intervals of all observational studies, with no falsification procedure.

4.3 Results

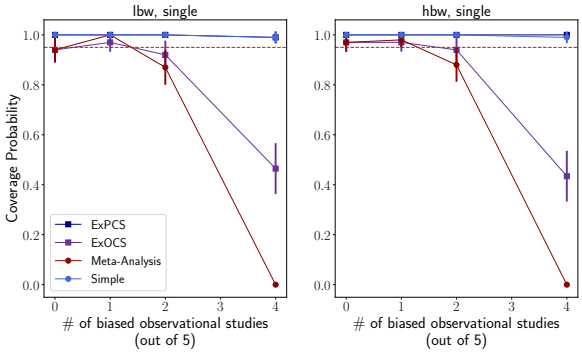
We perform three semi-synthetic experiments to assess the performance of our proposed meta-algorithm under different scenarios. The first experiment applies our algorithm under the default settings given in Section 4.1. In the second experiment, we vary the sample size ratio between the observational 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 4, 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. The goal is to produce confidence intervals that cover the



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.



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 4: Coverage probabilitis 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.

true GATEs in the extrapolated subgroups (i.e. (single, low birth weight) and (single, high birth weight)). Among the confidence intervals with adequate coverage, a narrower interval would then be preferred. 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 high-birth-weight, single-mother subgroup. This hints at the need for a conservative combination of non-falsified studies. However, an approach that is too conservative, such as *Simple Union*, produces intervals that are too wide and 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 *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 F, Table 1). 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 3show 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.

An analysis of including biased observational studies: In Figure 4, 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.

5 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 [23] 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 [18, 61, 1, 20], down-weighting studies with higher risk of bias [25, 34], or using Bayesian hierarchical regression to account for difference between subgroups of studies [43, 60]. 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) [17, 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 [46, 47, 62]. 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 G.

6 Discussion and Limitations

We have presented a meta-algorithm for constructing conservative confidence intervals for group average treatment effects for 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, with respect to the downstream societal impact of our work, great care should be taken to appropriately validate and soundly interpret the results of our method, especially with more sensitive subgroups (e.g. with respect to race or gender).

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

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

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

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

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

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

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

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

[60] 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.

[61] 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.

[62] 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 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (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]