Prediction-powered Generalization of Causal Inferences

Ilker Demirel [1 2] **Ahmed Alaa** [3] **Anthony Philippakis** [2] **David Sontag** [1]

Abstract

Causal inferences from a randomized controlled trial (RCT) may not pertain to a *target* population where some effect modifiers have a different distribution. Prior work studies *generalizing* the results of a trial to a target population with no outcome but covariate data available. We show how the limited size of trials makes generalization a statistically infeasible task, as it requires estimating complex nuisance functions. We develop generalization algorithms that supplement the trial data with a prediction model learned from an additional *observational* study (OS), without making *any* assumptions on the OS. We theoretically and empirically show that our methods facilitate better generalization when the OS is "high-quality", and remain robust when it is not, and *e.g.*, have unmeasured confounding.

1. Introduction

Experimental data from randomized controlled trials (RCT) is the gold standard for causal inference as various biases are avoided by design (Imbens & Rubin, 2015; Hernan & Robins, 2021). However, in addition to being time and cost-intensive, RCTs often exhibit limited external validity, and their findings may not apply to a *target population* (Rothwell, 2005; Stuart et al., 2011). The generalizability of an RCT is compromised when baseline factors that influence prognosis (*effect modifiers*) have different distributions in the trial and target populations (Dahabreh et al., 2019) (see Figure 1). For instance, trials may consist of healthier individuals on average than routine clinical practice. Since the overall health status likely affects the prognosis, it leads to "confounding" bias between the population-level effects in the trial and the target populations (Hernán et al., 2004).

[1]MIT CSAIL [2]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard [3]Department of Computational Precision Health, UC Berkeley and UCSF. Correspondence to: Ilker Demirel <demirel@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Dahabreh et al. (2019; 2020) develop methods that use individual-level covariate, treatment, and outcome data from a trial and only covariate information from the target population to estimate causal quantities in the latter (generalization). In this work, we show how combining trial data with potentially biased observational data, *e.g.*, found from electronic health records, can power better generalization.

Our Contributions We derive the generalization mean-squared error when an outcome model learned from the trial is used in the target sample to estimate an average causal effect in the target population, and probe how it increases when the trial is small and not representative of the target population (Section 3). Drawing inspiration from recent advances in using black-box models for valid statistical inference (Schuler et al., 2021; Angelopoulos et al., 2023), we develop *prediction-powered* estimators that leverage additional observational data without *any* assumptions on it and discuss when they lead to lower generalization error (Section 4). We simulate over a thousand data-generating processes and find that our estimators yield remarkable improvements when the observational data is high-quality and maintain baseline performance when it is not (Section 5).

Related Work There is growing interest in integrating data from trials and observational studies (OS) (Bareinboim & Pearl, 2016; Yang & Ding, 2019; NICE, 2022; Colnet et al., 2024). Schuler et al. (2021); Liao et al. (2023) show how adjustment by the predictions of a model learned from an OS can increase power in analyzing a trial. Similarly, Guo et al. (2021) investigate how coupling trial data and with "control-variates" constructed in an OS may enable smaller-variance estimation of the average treatment effect (ATE) in the *trial population*. Hartman et al. (2015); Degtiar et al. (2023) study generalization to a target population defined by the OS population or its union with the trial population. Han et al. (2023) study ATE estimation in a target population by incorporating data from multiple *source* populations, where the ATE is identifiable in *all* of the populations but different. Oberst et al. (2023) review methods that combine ATE estimates from a trial and an OS to obtain a better hybrid estimate (Rosenman et al., 2020; Cheng & Cai, 2021; Yang et al., 2023). Kallus et al. (2018); Chen et al. (2021); Hatt et al. (2022) consider the heterogeneity in effects and focus on the conditional ATE (CATE) *func-*

tion. Rosenman & Owen (2021) adopt a different angle and studies more efficient trial design using data from OS.

Another line of papers studies *benchmarking* evidence from OS (Forbes & Dahabreh, 2020). Hussain et al. (2022; 2023); Demirel et al. (2024) develop falsification tests for the causal assumptions by comparing the findings of an OS and a trial. De Bartolomeis et al. (2024a;b) focus on *quantifying* the hidden confounding in an OS, and Karlsson & Krijthe (2024) show how one can *detect* hidden confounding using multiple OS with a shared data-generating process.

In the works above, the target population of interest is taken as either the OS population or its union with the trial population. We consider a more general setting where the target population is defined separately from the trial and OS populations so long as it consists of trial-eligible individuals. For instance, the target population can represent a subgroup in the trial with a small sample size.

Our goal is to estimate population-level causal effects in the target population, for which only covariate information is available, by integrating data from a small trial and a large OS. We detail the necessary causal identification assumptions in Section 2, which crucially do not enforce any unverifiable conditions on the OS, and describe our estimators in Section 4. We do not go the route of cooking a recipe to combine real-valued estimates from the trial and the OS, nor promise to give guarantees on the granular CATE function, as the former offers poor flexibility in utilizing rich observational data and the latter replaces the *causal* assumptions on the OS with statistical assumptions on its bias function. Our approach lies somewhere in between: we fit an outcome *function* from the OS using flexible machine learning models, which can be subject to *causal biases*, and analyze how coupling it with trial data can power the estimation of a *real-valued* causal estimand in the target population.

2. Background

2.1. Notation and Objective

We consider a *nested* design where a trial is sampled from an underlying population of trial-eligible individuals. Note that our methods can easily be extended to *nonnested* designs where the target sample is obtained separately; *e.g.*, to represent a subgroup for which the trial sample alone cannot power statistically significant inference.

We have access to an i.i.d. sample of observations $\mathcal{D} = \{W_i\}_{i=1}^n$ with $W_i = (X_i, S_i, S_i \times A_i, S_i \times Y_i)$, where $X \in \mathcal{X}$ is a d-dimensional covariate vector, S is a binary trial participation indicator, A is a categorical treatment, and $Y \in \mathbb{R}$ is the outcome of interest. Only covariate data is available for non-participants ($S = 0$), while treatment and outcome data are also available for participants ($S = 1$).

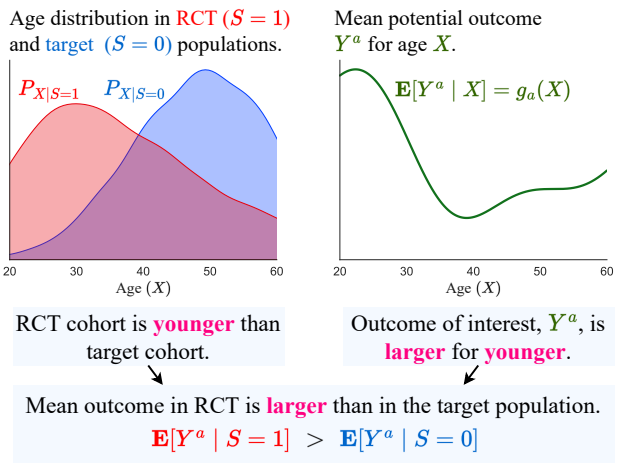

Figure 1. Age influences both selection into the trial and the outcome, inducing confounding bias between the population-level mean potential outcomes in the trial and target populations.

The *target* population of interest is represented by non-participants. We denote by $\mathcal{D}_1 \subset \mathcal{D}$ the set of trial participants and by $\mathcal{D}_0 \subset \mathcal{D}$ the set of non-participants with sizes $n_1 = \sum_{i=1}^n \mathbf{1}\{S_i = 1\}$ and $n_0 = \sum_{i=1}^n \mathbf{1}\{S_i = 0\}$, partitioning the composite sample $\mathcal{D}$. Further, we denote by P_1 and P_0 the joint distribution of W in the underlying trial and target populations. For instance, $X \sim P_0$ represents a covariate drawn from the target distribution $P(X \mid S = 0)$.

We seek causal inference in the target population. Specifically, denoting by Y^a the *potential* outcome under treatment $A = a$, we want to estimate the average potential outcomes in the target population.

$$\mu_a := \mathbf{E}[Y^a \mid S = 0]. \tag{1}$$

A more common causal estimand, the average treatment effect, can be directly derived from the average potential outcomes (*e.g.*, $\mu_1 - \mu_0$ in a binary treatment setting). Focusing on potential outcomes allows for simpler exposition, and they are of independent interest in many applications.

The challenge in estimating μ_a is three-fold. The first is obvious: no outcome data is available for non-participants. One can contemplate resorting to outcome data from the trial, which brings us to the second challenge. The potential outcome Y^a can only be observed for those who received treatment $A = a$. When treatment assignment depends on *unobserved* factors that also affect the outcome, one risks *confounding* bias, which presents a non-trivial challenge in analyzing *observational* data (see Section 4). However, it is easily avoided in trials by *randomized* treatment assignment, and the average potential outcome in *trial population*, $\mathbf{E}[Y^a|S = 1]$, can be reliably estimated. The final challenge

is $\mathbf{E}[Y^a|S=0] \neq \mathbf{E}[Y^a|S=1]$ when there is confounding by *trial participation*, leading to different distributions of effect modifiers in trial and target populations (see Figure 1). Therefore, one cannot generalize population-level effect estimates from a trial to the target population, but needs to adjust for confounding by trial participation.

Next, we state the causal assumptions needed to estimate μ_a by incorporating outcome data from the trial.

2.2. Assumptions for Causal Inference

Assumption 2.1 (*Consistency*). $A = a \implies Y = Y^a$.

Assumption 2.2 (*Mean ignorability of treatment assignment in trial*). $\mathbf{E}[Y^a \mid X, S=1] = \mathbf{E}[Y^a \mid X, S=1, A=a]$.

Assumption 2.3 (*Positivity of treatment assignment in trial*). $P(A=a \mid X=x, S=1) > 0$.

Assumptions 2.1-2.3 are satisfied in an RCT by design, and they enable causal inference within the trial population, *i.e.*, reliable estimation of $\mathbf{E}[Y^a \mid S=1]$.

Assumption 2.4 (*Mean ignorability of trial participation*). $\mathbf{E}[Y^a \mid X] = \mathbf{E}[Y^a \mid X, S=1] = \mathbf{E}[Y^a \mid X, S=0]$.

Assumption 2.5 (*Positivity of selection into trial*). $P(S=0 \mid X=x) > 0 \implies P(S=1 \mid X=x) > 0$.

Assumption 2.4 requires that within levels of measured covariates X, potential outcomes in the trial and target populations are the same on average. Assumption 2.5 ensures that every patient has a nonzero probability of participating in the trial, and we do not have to rely on pure *extrapolation*. Assumptions 2.4 and 2.5 transform the problem of "generalizing the results of a trial" into a *covariate* shift problem and allow one to identify μ_a as follows (Dahabreh et al., 2020).

$$\begin{aligned}\mu_a &= \mathbf{E}_{X\sim P_0}[\mathbf{E}[Y^a \mid X, S=0]] \\ &= \mathbf{E}_{X\sim P_0}[\mathbf{E}[Y^a \mid X, S=1]] \\ &= \mathbf{E}_{X\sim P_0}[\mathbf{E}[Y \mid X, S=1, A=a]]. \qquad (2)\end{aligned}$$

where last two steps follow from Assumptions 2.4-2.5 and 2.1-2.3. Note that (2) can be estimated using only covariate data from non-participants ($S=0$) and covariate, treatment, and outcome data from the trial participants ($S=1$).

3. Generalization Using Experimental Data

Dahabreh et al. (2020) propose estimators of (2) based on outcome functions, weighting by the inverse of the trial participation probability, and doubly-robust (DR) ones. We focus on the outcome function approach as it more lucidly uncovers the limitations of generalization from trial data, and the synthetic results in Dahabreh et al. (2020) show that it outperforms the weighting-based approaches and performs on par with the DR ones (as we also verify in Appendix B.2). Our findings reveal how a predictive model trained on large-scale observational data could help.

We define the mean outcome function in the trial population $S=1$ under treatment $A=a$ as

$$\begin{aligned}g_a(X) &:= \mathbf{E}[Y \mid X, S=1, A=a] \qquad (3)\\ &= \mathbf{E}[Y^a \mid X, S=1] \qquad \text{(Assumptions 2.1-2.3)}\\ &= \mathbf{E}[Y^a \mid X]. \qquad \text{(Assumptions 2.4-2.5)}\end{aligned}$$

One can estimate $\hat{g}_a(X)$ from the trial sample $\mathcal{D}_1$, and then average its predictions in the target sample $\mathcal{D}_0$. This leads to the following outcome model (OM) estimator on the composite sample $\mathcal{D}$.

$$\hat{\mu}_a^{\text{OM}} = \frac{1}{n_0}\sum_{i=1}^{n} \mathbf{1}\{S_i=0\}\,\hat{g}_a(X_i). \qquad (4)$$

In the remainder of this section, we investigate when $\hat{\mu}_a^{\text{OM}}$ is expected to have high mean-squared error (MSE). Our next result gives an approximation for the MSE in the special case where X is purely categorical, which provides perspective into the limitations of $\hat{\mu}_a^{\text{OM}}$ for the more general case as well.

Proposition 3.1. *Let X be a categorical covariate stratifying the population into K groups and denote by $n_{s=1,a,k}$ the number of trial participants from group $X=k$ assigned to treatment $A=a$, and by $\sigma^2_{a,k}$ the variance of outcome among such patients. Let us estimate the outcome function $g_a(X=k)$ with the sample mean of outcomes Y of participants in group $X=k$ assigned to treatment $A=a$.*

Suppose that Assumptions 2.1-2.5 hold. When n_0 is large, the MSE of $\hat{\mu}_a^{\text{OM}}$ in (4) can be approximated as

$$\mathbf{E}[(\hat{\mu}_a^{\text{OM}} - \mu_a)^2] \approx \sum_{k=1}^{K} p^2_{s=0}(k)\frac{\sigma^2_{a,k}}{n_{s=1,a,k}}, \qquad (5)$$

where $p_{s=0}(k) := P(X=k \mid S=0)$ is the proportion of patients from group $X=k$ in the target population.

Proposition 3.1 reveals the key challenge in our endeavor. The reason behind the need for a "generalization procedure" is that some effect modifiers' distributions might differ in the trial and target populations. Reading off (5), one can see that when the trial is limited in representing patient profiles that are prevalent in the target population (small $n_{s=1,a,k}$, large $p_{s=0}(k)$), the MSE will be larger. That is, inference in target population gets more challenging when it becomes "more different" from the trial population.

The insights from Proposition 3.1 extend to the case with continuous covariates and parametric estimators $\hat{g}_a(X) = g_a(X;\hat{\theta})$ (*e.g.*, a random forest). Let us denote by $\mathcal{A}$ the algorithm that fits $\hat{\theta}$ from the trial sample (*e.g.*, ridge regression), *i.e.*, $\hat{\theta} = \mathcal{A}(\mathcal{D}_1)$. As $\mathcal{D}_1 \sim P_1$, we write $\hat{\theta} \sim \mathcal{A}(P_1)$ to refer to the randomness in estimating $\hat{\theta}$ from $\mathcal{D}_1$. Next, we give an approximation for the MSE in the general case.

Theorem 3.2. *Suppose that Assumptions 2.1-2.5 hold and consider a parametric estimator $\hat{g}_a(X) = g_a(X;\hat{\theta})$ for the outcome function. For large n_0, the MSE of $\hat{\mu}_a^{\text{OM}}$ in (4) can be approximated as*

$$\mathbf{E}[(\hat{\mu}_a^{\text{OM}} - \mu_a)^2]$$
$$\approx \mathbf{E}_{X\sim P_0}\Big[\underbrace{\mathbf{E}_{\hat{\theta}\sim\mathcal{A}(P_1)}[g_a(X;\hat{\theta})] - g_a(X)}_{=:\,\text{SB}_g(X)}\Big]^2 \quad (6)$$
$$+ \text{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\big(\mathbf{E}_{X\sim P_0}[g_a(X;\hat{\theta})]\big). \quad (7)$$

The first term, (6), is the statistical bias (SB). Crucially, it is obtained by integrating the bias function $\text{SB}_g(X)$ over the *target* covariate distribution $P_{X|S=0}$, making $\hat{\mu}_a^{\text{OM}}$ susceptible to weak overlap between trial and target populations. Consider the case where $g_a(X;\hat{\theta})$ is misspecified/underfit. $\text{SB}_g(X)$ will be larger where $P_{X|S=1}$ has little weight, since $g_a(X;\hat{\theta})$ is fit using the *trial* sample $\mathcal{D}_1$. $\hat{\mu}_a^{\text{OM}}$ may then suffer substantial bias if $P_{X|S=0}$ is large in covariate regions where the trial support is weak. It is therefore essential to ensure that $g_a(X;\hat{\theta})$ is rich enough and can match the complexity of $g_a(X)$ to avoid a large bias term. However, in practice, one's ability to flexibly model $g_a(X;\hat{\theta})$ is severely limited as the small size of trials (*e.g.,* $\sim$ 200) can lead to overfitting, *i.e.*, increasing the variance term (7). We empirically demonstrate this tradeoff in Appendix B.1.

4. Prediction-powered Generalization Using Experimental and Observational Data

Here, we study how integrating rich observational data with limited experimental data can make the generalization task more statistically feasible. We index by $S = 2$ the observational population with joint distribution P_2. We assume access to an i.i.d. sample of observations $(X_i, A_i, Y_i) \sim P_2$ and denote by $\mathcal{D}_{2,a}$ the set of patients who received treatment $A = a$ in the observational data. In the first step, we fit a predictive model $f_a : \mathcal{X} \to \mathbb{R}$ by minimizing the empirical mean-squared error in $\mathcal{D}_{2,a}$ to approximate

$$\mathbf{E}[Y \mid X, S = 2, A = a]. \quad (8)$$

Unlike the trial sample, large observational data can support parametrizing $f_a(X)$ with powerful machine learning models, allowing it to model complex functions.

If one is willing to make Assumptions 2.1-2.5 for the observational study (OS), *i.e.*, $S = 2$ instead of $S = 1$, μ_a can be identified as $\mathbf{E}_{X\sim P_0}[\mathbf{E}[Y|X, S = 2, A = a]]$ via the same machinery in (2). One could then apply $f_a(X)$ in the composite sample $\mathcal{D}$ to estimate μ_a as

$$\hat{\mu}_a^{\text{OS-OM}} = \frac{1}{n_0}\sum_{i=1}^{n} \mathbf{1}\{S_i = 0\}\, f_a(X_i), \quad (9)$$

analogous to (4). While Assumptions 2.1-2.5 are defensible for trials, some of them rarely hold for observational studies in practice, particularly the ignorability of treatment assignment (no unmeasured confounding). In contrast to most of the literature on causal inference using observational data, we take an extremely assumption-light approach, making *no* assumptions on observational data. We define the "bias function" as the difference between the outcome function in the trial, $g_a(X)$, and the observational predictor, $f_a(X)$.

$$\begin{aligned} b_a(X) &:= f_a(X) - g_a(X) \quad (10)\\ &= \underbrace{f_a(X) - \mathbf{E}[Y \mid X, S = 2, A = a]}_{\text{statistical bias}} \\ &\quad + \underbrace{\mathbf{E}[Y^a \mid X, S = 2, A = a] - \mathbf{E}[Y^a \mid X, S = 2]}_{\text{confounding bias}} \\ &\quad + \underbrace{\mathbf{E}[Y^a \mid X, S = 2] - \mathbf{E}[Y^a \mid X, S = 1]}_{\text{transportation bias}}, \end{aligned}$$

since $\mathbf{E}[Y^a \mid X, S = 1] = g_a(X)$ (see (3)). The statistical bias term is related to fitting $f_a(X)$ using a finite sample, and it vanishes with more data given enough model capacity. Confounding and transportation biases, however, are the price of avoiding Assumptions 2.2 and 2.4 for the observational study ($S = 2$). They will not disappear even with infinite data from the observational population, rendering $\hat{\mu}_a^{\text{OS-OM}}$ in (9) an *inconsistent* estimator for $\mu_a = \mathbf{E}[Y^a \mid S = 0]$ even when $f_a(X) = \mathbf{E}[Y \mid X, S = 2, A = a]$.

In Sections 4.1 and 4.2, we derive two new identifications of μ_a that integrate the predictions of f_a in a statistically valid way and discuss how they lead to more sample-efficient estimation in comparison to (2). We give regression function-based estimators, derive their MSEs, and compare them to that of the baseline in Theorem 3.2.

4.1. Additive Bias Correction to Predictive Model

We covered why using f_a alone is unreliable. Nonetheless, it may carry useful signal we can exploit when coupled with trial data. First, using the trial sample $\mathcal{D}_1$, we show how one can learn the bias function of the predictive model, $b_a(X) = f_a(X) - g_a(X)$. We then give an estimator for μ_a that uses the predictions $f_a(X)$ in the target sample by correcting with their estimated bias, $\hat{b}_a(X)$. We formalize in Theorem 4.2 and Section 4.1.3 when it is more advantageous to construct an estimator of μ_a that relies on fitting the bias function $b_a(X)$ instead of the outcome function $g_a(X)$, such as the illustrative example depicted in Figure 2.

4.1.1. Identification

We start by trivially writing

$$\mu_a = \mathbf{E}\left[f_a(X) \mid S = 0\right] - \mathbf{E}\left[f_a(X) - Y^a \mid S = 0\right]. \quad (11)$$

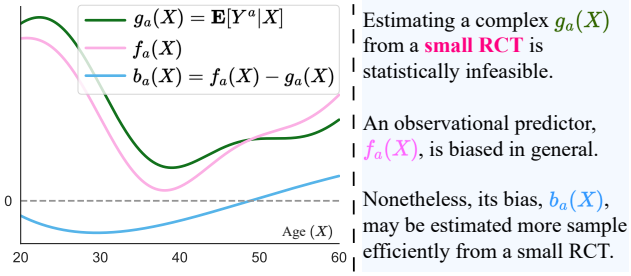

Figure 2. A biased predictor $f_a(X)$ can still capture higher order polynomials, making its bias $b_a(X)$ "easier" to learn than $g_a(X)$.

The first term can be estimated by averaging $f_a(X)$ in the target sample, which is generally biased for μ_a. The second term "removes" this bias; however, as it contains a *counterfactual* variable, Y^a, it is not immediately clear how one would estimate it. Our next result shows that it can be identified without additional assumptions on f_a.

Lemma 4.1. *Suppose that Assumptions 2.1-2.5 hold. Let $f_a : \mathcal{X} \to \mathbb{R}$ and define the error variable*

$$Z \coloneqq f_a(X) - Y. \tag{12}$$

μ^a can be identified as

$$\mu_a = \mathbf{E}_{X\sim P_0}[f_a(X) - \mathbf{E}[Z \mid X, S = 1, A = a]]. \tag{13}$$

Note that Z can be calculated for trial participants and second term in (13) can be estimated with covariate information from the target sample and covariate, treatment, and "error" information from the trial sample, as we cover next.

4.1.2. Regression Function-based Estimation

By (10), (12), and (3), it is straightforward to see that

$$b_a(X) = \mathbf{E}[Z \mid X, S = 1, A = a].$$

That is, the identification in (13) is through the bias function in (10). We denote by $b_a(X;\hat{\gamma})$ a parametric fit obtained by regressing Z onto covariates X in the trial sample and write the additive-bias-correction (ABC) estimator.

$$\hat{\mu}_a^{\text{ABC}} = \frac{1}{n_0}\sum_{i=1}^{n} \mathbf{1}\{S_i = 0\}\left(f_a(X_i) - b_a(X_i;\hat{\gamma})\right). \tag{14}$$

Theorem 4.2. *Suppose that Assumptions 2.1-2.5 hold. For large n_0, the MSE of $\hat{\mu}_a^{\text{ABC}}$ in (14) can be approximated as*

$$\begin{aligned}&\mathbf{E}[(\hat{\mu}_a^{\text{ABC}} - \mu_a)^2]\\ &\approx \mathbf{E}_{X\sim P_0}\left[\mathbf{E}_{\hat{\gamma}\sim\mathcal{A}(P_1)}[b_a(X;\hat{\gamma})] - b_a(X)\right]^2 &(15)\\ &\quad + \mathrm{Var}_{\hat{\gamma}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}[b_a(X;\hat{\gamma})]\right). &(16)\end{aligned}$$

Algorithm 1 Generalization via additive bias correction

Input: Sample $\mathcal{D}$, Predictor f_a, MSE optimizer $\mathcal{A}$
$\mathcal{D}_{1,a} \subset \mathcal{D}$: Trial cohort ($S_i = 1$) with treatment $A_i = a$
for $W_i \in \mathcal{D}_{1,a}$ **do**
  Calculate the prediction error $Z_i = f_a(X_i) - Y_i$
end for
Fit $b_a(X;\hat{\gamma})$ by minimizing MSE for Z in $\mathcal{D}_{1,a}$ using $\mathcal{A}$
Return $\hat{\mu}_a^{\text{ABC}}$ in (14)

The significance of Theorem 3.2 is showing that the MSE of $\hat{\mu}_a^{\text{ABC}}$ admits the same form with that of $\hat{\mu}_a^{\text{OM}}$ in Theorem 3.2. The difference is that the MSE is governed by how well the bias function $b_a(X)$ is estimated instead of the outcome function $g_a(X)$. This result formalizes how leveraging a potentially biased observational predictor can be more viable for the "generalization" task. Consider the case in Figure 2 where $f_a(X)$ captures higher degree polynomials in $g_a(X)$, resulting in $b_a(X)$ being a low-degree polynomial. One can then fit a linear model with a few polynomial features for $b_a(X;\hat{\gamma})$, resulting in both controlled bias (15) and variance (16) terms. On the other hand, fitting $g_a(X;\hat{\theta})$ similarly will result in a large bias term (6). We provide a detailed discussion in the next section and empirically demonstrate how the bias ((6), (15)) and variance ((7), (16)) terms compare in Appendix B.1.

4.1.3. Case Study: Polynomial Ridge Regression

The symmetry between the MSEs in Theorems 3.2 and 4.2 allows one to reason about the (comparative) performances of the outcome and bias function-based estimators in (4) and (14). To gain further insight, we study the polynomial ridge regression framework and describe the regime where estimating $b_a(X)$ is more feasible than $g_a(X)$ in the setting described below. We focus on *finite-sample* results, which are of significant interest given the limited size of trials.

We consider $X \in [-1,1]$, denote by $L^2([-1,1])$ the space of square-integrable functions [1] endowed with the inner-product $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x$, and assume that $g_a, b_a \in L^2([-1,1])$ with bounded norms $\|g_a\|, \|b_a\| \le 1$. Finally, we assume the following generative equations.

$$Y_i = g_a(X_i) + \eta_i, \qquad Z_i = b_a(X_i) - \eta_i, \tag{17}$$

where $\eta_i \sim \mathcal{N}(0,\sigma^2)$ are zero-mean i.i.d. noise variables and Z_i are the patient-wise *error* terms for the predictive model $f_a(X)$, defined previously in (12).

Let us now define, for a generic function f, the "empirical excess risk" of a fit $\hat{f}$ obtained from a sample of size m as

$$R_m(\hat{f}, f) \coloneqq \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(X_i) - f(X_i))^2, \tag{18}$$

[1] $\int_{-1}^{1} f^2(x)\,\mathrm{d}x < \infty, \quad \forall f \in L^2([-1,1])$.

which quantifies how far away the fit is from the true function. In the rest of this section, we study oracle upper bounds on $R_{n_1}(\hat{g}_a, g_a)$ and $R_{n_1}(\hat{b}_a, b_a)$ when $\hat{g}_a$ and $\hat{b}_a$ are fit via polynomial ridge regression in the trial sample $\mathcal{D}_1$. To that end, let us now introduce the Legendre polynomials which have convenient properties that facilitate clear exposition. We denote by $\phi_k : [-1, 1] \to \mathbb{R}$ the k-th order *normalized* Legendre polynomial. The set $\{\phi_k\}_{k=0}^{\infty}$ form an *orthonormal basis* [2] for $L^2([-1, 1])$; meaning that any function $f \in L^2([-1, 1])$ can be uniquely represented as a linear combination of $\{\phi_k\}_{k=0}^{\infty}$, allowing us to write

$$g_a(X) = \sum_{k=0}^{\infty} \lambda_k \phi_k(X), \quad b_a(X) = \sum_{k=0}^{\infty} \omega_k \phi_k(X), \quad (19)$$

where $\lambda_k = \langle g_a, \phi_k \rangle$ and $\omega_k = \langle b_a, \phi_k \rangle$. In practice, one can fit $\hat{g}_a$ and $\hat{b}_a$ using Legendre polynomials up to degree $d' \in \mathbb{N}$ with ridge regularization to avoid overfitting to the trial sample. Leaving the intermediary steps to Appendix A.2, we proceed to state the corresponding upper bounds on the expected empirical excess risks.

Lemma 4.3 (Adopted from Wainwright (2019))**.** *Let $X \in [-1, 1]$, $g_a, b_a \in L^2([-1, 1])$, $\|g_a\|, \|b_a\| \leq 1$, and consider the generative equations in (17) with noise variance σ^2. Denote by $\hat{g}_a$ and $\hat{b}_a$ the fits obtained by regressing Y and Z (see (12)) onto $\{\phi_k(X)\}_{k=0}^{d'}$ in trial sample $\mathcal{D}_1$ with an appropriately chosen ridge regularization penalty. We have*

$$\mathbf{E}_{X_i, \eta_i}[R_{n_1}(\hat{g}_a, g_a)] \leq \sigma^2 d'/n_1 + \sum\nolimits_{k=d'+1}^{\infty} \lambda_k^2, \quad (20)$$

$$\mathbf{E}_{X_i, \eta_i}[R_{n_1}(\hat{b}_a, b_a)] \leq \sigma^2 d'/n_1 + \sum\nolimits_{k=d'+1}^{\infty} \omega_k^2. \quad (21)$$

The upper bounds in (20) and (21) share the first statistical error term, which grows with the number of polynomial features d'. Comparing the second terms reveals that estimating the bias function is favorable when $\sum_{k=d'+1}^{\infty} \omega_k^2 < \sum_{k=d'+1}^{\infty} \lambda_k^2$. One would expect the preceding condition to hold in two scenarios, which we discuss next.

The first one is when the observational predictor is high quality. Precisely, if $f_a(X) \approx g_a(X)$, then $b_a(X) \approx 0$, implying small values for ω_k and sum of their squares. This is the same condition in Angelopoulos et al. (2023) for a black-box predictor to power better inference when coupled with a small amount of gold-standard data. In the context of *causal* inference, it would take the individual terms in (10) to be as small as possible to warrant $f_a(X) \approx g_a(X)$, which requires observational study to have negligible hidden confounding for treatment assignment and to be transportable conditioned on X.

The second scenario is when b_a "mostly" consists of lower degree polynomials, that is, $w_k \approx 0$ for $k > d'$. This is a relaxed and more general version of the key assumption in Kallus et al. (2018), which requires $b_a(X)$ to be linear in X. The idea is that even when $f_a(X)$ is biased, it can still capture complex structure, such as the higher order polynomials modeling rapid turns in $g_a(X)$, and make $b_a(X)$ considerably simpler, as illustrated in Figure 2.

[2] $\langle \phi_i, \phi_j \rangle = \delta_{ij}$, $\text{span}(\{\phi_k\}_{k=0}^{\infty}) = L^2([-1, 1])$.

4.2. Augmented Outcome Modeling

Here we draw from Schuler et al. (2021); Liao et al. (2023) and leverage the observational model by using its predictions as an additional regressor while estimating the outcome function from the trial.

Using $f_a(X)$ as a covariate still makes for an easier estimation task when $g_a(X) = f_a(X) + b_a(X)$ with $f_a(X)$ capturing most of the complexity in $g_a(X)$ and $b_a(X)$ is a simpler function. However, it has two advantages over the additive bias correction approach we discuss below.

First is **robustness** when $f_a(X)$ does not carry useful information. For instance, let $f_a(X)$ be an independent noise term, η, for all X. Then the additive bias $b_a(X) = \eta - g_a(X)$ is just a noisier version of $g_a(X)$ and even more challenging to estimate. On the other hand, when the predictions $f_a(X)$ are used as a covariate, a good learning algorithm would just ignore it. We compare the two approaches' robustness with synthetic experiments (see Figure 5).

Second is the **flexibility** in how the predictions are utilized. Consider the illustrative example where $g_a(X) = f_a(X)/2$ and the bias of $f_a(X)$ can be corrected simply dividing it by two. On the other hand, additive bias $b_a(X) = g_a(X)$ is identical to the outcome function and not easier to estimate.

4.2.1. Identification

Let us define the *augmented* covariate vector as

$$\tilde{X}_i \coloneqq [X_i^1, X_i^2, \ldots, X_i^d, f_a(X_i)], \quad (22)$$

where X_i^n is the n-th *original* covariate out of d. We denote $\tilde{X} \in \tilde{\mathcal{X}}$ where $\tilde{\mathcal{X}} = \mathcal{X} \times \mathbb{R}$.

Lemma 4.4. *Suppose that Assumptions 2.1-2.5 hold. Let $f_a : \mathcal{X} \to \mathbb{R}$ and define the augmented outcome function*

$$h_a(\tilde{X}) \coloneqq \mathbf{E}[Y \mid \tilde{X}, S = 1, A = a]. \quad (23)$$

where $\tilde{X}$ is defined in (22). μ^a can be identified as

$$\mu_a = \mathbf{E}_{X \sim P_0}[h_a(\tilde{X})]. \quad (24)$$

Note that X and $\tilde{X}$ carry the same *information* and Assumptions 2.1-2.5 continue to hold for $\tilde{X}$. The identification in (24) thus follows from the same steps that lead to (2).

Algorithm 2 Generalization via augmented outcome model

Input: Sample $\mathcal{D}$, Predictor f_a, MSE optimizer $\mathcal{A}$
$\mathcal{D}_{1,a} \subset \mathcal{D}$: Trial cohort ($S_i = 1$) with treatment $A_i = a$
for $W_i \in \mathcal{D}$ **do**
  Calculate the outcome prediction $f_a(X_i)$
  Construct the augmented covariate vector $\tilde{X}_i$ as in (22)
end for
Fit $h_a(\tilde{X}; \hat{\beta})$ by minimizing MSE for Y in $\mathcal{D}_{1,a}$ using $\mathcal{A}$
Return $\hat{\mu}_a^{\text{AOM}}$ in (25)

4.2.2. Regression Function-based Estimation

Note that $h_a(\tilde{X}) = h_a(X, f_a(X)) = g_a(X)$ and the only difference from the baseline approach in Section 3 is that we have an additional regressor, $\tilde{X}_{d+1} = f_a(X)$. We denote by $h_a(\tilde{X}_i; \hat{\beta})$ the parametric fit for $h_a(\tilde{X}_i)$ and write the *augmented* outcome modeling (AOM) estimator as

$$\hat{\mu}_a^{\text{AOM}} = \frac{1}{n_0} \sum_{i=1}^{n} \mathbf{1}\{S_i = 0\}\, h_a(\tilde{X}_i; \hat{\beta}). \tag{25}$$

The approximation to the MSE $\mathbf{E}[(\hat{\mu}_a^{\text{AOM}} - \mu_a)^2]$ follows the same form with that of $\hat{\mu}_a^{\text{OM}}$ in Theorem 3.2 with the augmented outcome function h_a replacing g_a. In the interest of space, we defer the precise statement to Appendix A.3.

We close this section by mentioning two critical directions for future work to leverage observational data more efficiently: one related to *modeling* and the other to *estimation*.

Representation-powered Outcome Modeling Instead of using a model's predictions $f_a(X)$, one can use the representations learned by the model as additional covariates in the trial (Johansson et al., 2016; Shalit et al., 2017). This approach also allows for extracting richer information from the observational data more flexibly, *e.g.*, via unsupervised and multimodal learning methods.

Doubly-robust Estimation For the prediction-powered identifications of μ_a in (13) and (24), we focused only on regression function-based estimators to demonstrate the advantages of our approach. In Appendix A.4, we give doubly-robust estimators that enjoy desirable properties such as asymptotic normality that enable the construction of confidence intervals (Chernozhukov et al., 2018; Kennedy, 2023).

5. Synthetic Experiments

We simulate over a thousand different synthetic data generating processes with varying levels of complexity in the outcome function $g_a(X)$, confounding bias in the observational study, and trial size n_1. We compare the root MSE (RMSE)

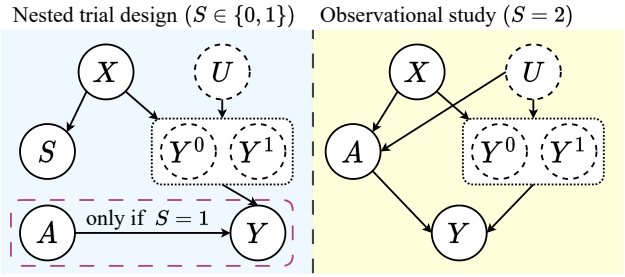

Figure 3. Data-generating process used in simulated experiments. (*Left.*) X (observed) induces confounding by trial participation. (*Right.*) In the observational study, there is *hidden* confounding for treatment assignment due to U (unobserved).

of our estimators (14) and (25), which combine experimental and observational data, to that of the baselines (4) and (9) which use them alone. Bias-variance terms (*e.g.,* (15) and (16)) are presented in Appendix B.1. Further, we demonstrate the robustness of the augmented outcome modeling estimator in (25) over the additive bias correction estimator in (14). While the main results are concerned with outcome-modeling-based estimators, we present additional empirical results for the inverse propensity weighting and and doubly-robust estimators in Appendix B.2. Our code is available at `https://github.com/demireal/ppci`.

5.1. Data-generating Process

We consider two covariates $X, U \in [-1, 1]$, a binary treatment strategy $A \in \{0, 1\}$, and a real-valued outcome $Y \in \mathbb{R}$. We first describe the probabilistic model that generates the potential outcomes Y^0 and Y^1 conditioned on X and U. We then move on to explain the sampling mechanism in the nested trial design that generates the trial and target cohorts. Finally, we specify the patient sampling and treatment assignment mechanism underpinning the observational data we use to train a predictive model $f_a : [-1, 1]^2 \to \mathbb{R}$. Results of simpler experiments where the functions underlying the data-generating process are specified to be linear are presented in Appendix B.4.

Generating (Potential) Outcomes We denote the *full* outcome model (FOM) for treatment $A = a$ by $\text{FOM}_a : [-1, 1]^2 \to \mathbb{R}$. For a patient with covariates (X_i, U_i), the potential outcome is calculated as $Y_i^a = \text{FOM}_a(X_i, U_i)$. Note that we use the same outcome model for patients in the trial, target, and observational samples.

We generate FOM_a by sampling from a GP with mean function $m(X, U) = 0$ and kernel function $k((X, U), (X', U'))$ (Rasmussen et al., 2006). We create a composite kernel by adding a squared-exponential (SE) kernel to model the *local* variations and a linear kernel to model the trends in the

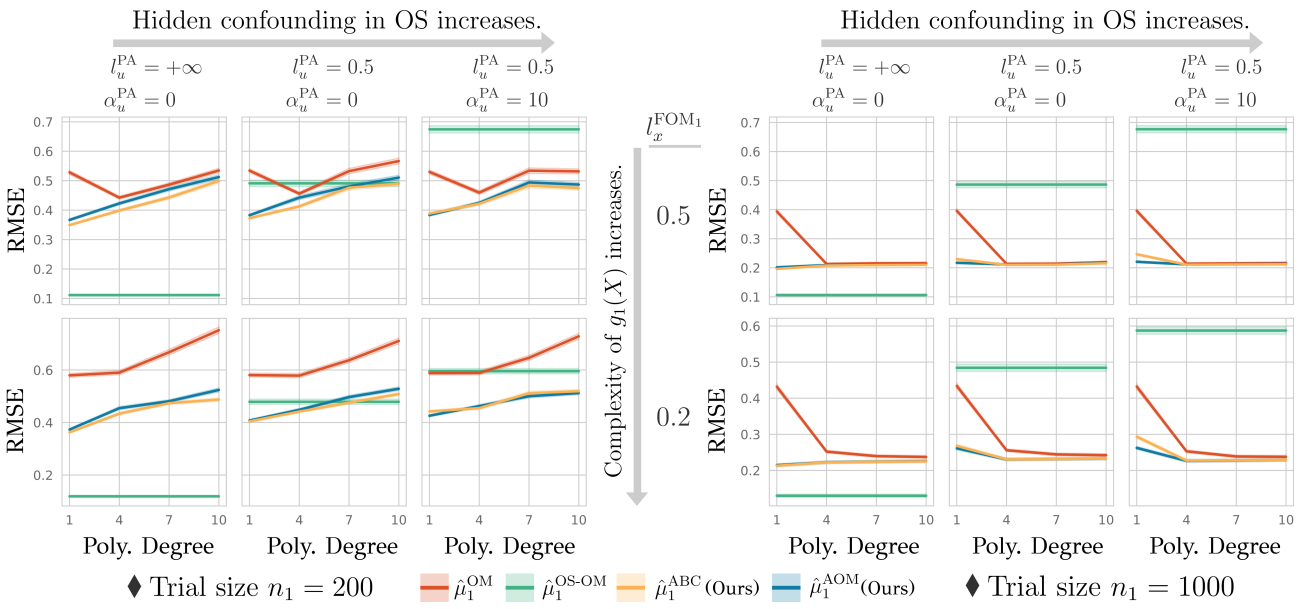

Figure 4. 100 different set of data-generating functions are sampled for each $(l_x^{\text{FOM}_1}, \alpha_u^{\text{PA}}, n_1)$. We plot the RMSE averaged over 100 scenarios. Results are reported for four different numbers of polynomial features used to fit the underlying regression functions (if any).

outcome. Precisely, we have

$$k((X,U),(X',U')) = \alpha_x^{\text{FOM}_a} XX' + \alpha_u^{\text{FOM}_a} UU' \quad \text{(linear)}$$
$$+ \exp\left(-\frac{(X-X')^2}{2(l_x^{\text{FOM}_a})^2} - \frac{(U-U')^2}{2(l_u^{\text{FOM}_a})^2}\right), \qquad \text{(SE)} \quad (26)$$

where $\alpha_x^{\text{FOM}_a}, \alpha_u^{\text{FOM}_a}, l_x^{\text{FOM}_a}, l_u^{\text{FOM}_a} \in \mathbb{R}_+$ are free parameters. We experiment with different values to simulate a diverse set of scenarios. For instance, a larger value for $\alpha_u^{\text{FOM}_a}$ implies a stronger linear trend in $\text{FOM}_a(X,U)$ along U-axis. More details are given at the end of this section.

Generating Trial and Target Samples We consider a *nested* study design and generate a composite *trial-eligible* patient cohort by sampling $X_i, U_i \sim \texttt{Uniform}[-1,1]$ independently. We denote by $P(S=1 \mid X_i, U_i)$ the probability of trial participation, which is generated as

$$P(S=1|X_i,U_i) = \text{median}\left\{\frac{1}{1+e^{-\text{L}_{\text{PS}}(X_i,U_i)}}, 0.1, 0.9\right\}. \quad (27)$$

where the "logit" function $\text{L}_{\text{PS}}(X_i,U_i)$ is sampled from a GP with the composite linear + SE kernel in (26) with parameters $\alpha_x^{\text{PS}} = 10$, $l_x^{\text{PS}} = 1$, $\alpha_u^{\text{PS}} = 0$, $l_u^{\text{PS}} = +\infty$. The last two parameters effectively imply that the trial participation probability does not depend on U but X only, ensuring Assumption 2.4. Taking a median with 0.1 and 0.9 ensures Assumption 2.5. Trial participation is then sampled as $\texttt{Bernoulli}(P(S=1 \mid X_i, U_i))$.

Finally, for trial participants ($S_i = 1$), the treatment assignment is sampled as $A_i \sim \texttt{Bernoulli}(0.5)$ and the observed outcome is generated as $Y = \text{FOM}_{A_i}(X_i, U_i)$, which ensures that Assumptions 2.1-2.3 hold.

Generating an Observational Sample An observational cohort is generated by sampling $X_i, U_i \sim \text{Uniform}[-1,1]$ independently. For each patient, treatments are sampled as $A_i \sim \texttt{Bernoulli}(P(A=1|S=2,X_i,U_i))$, where the probability of treatment assignment is generated similar to (27) through a logit function $\text{L}_{\text{PA}}(X,U)$ sampled from a GP with parameters $\alpha_x^{\text{PA}}, \alpha_u^{\text{PA}}, l_x^{\text{PA}}, l_u^{\text{PA}} \in \mathbb{R}_+$. The observed outcomes are generated as $Y = \text{FOM}_{A_i}(X_i, U_i)$.

Simulated Scenarios and GP Parameters We focus on the mean potential outcome under treatment $A=1$ in the target population, $\mu_1 = \mathbf{E}[Y^1 \mid S=0]$. The sample size for the observational study (OS) and the target sample are set to $50,000$ and $20,000$, respectively. We experiment with different values for the parameters $n_1, l_x^{\text{FOM}_1}$, and α_u^{PA}. We fit $f_1(X)$ from the OS with a neural network, and $g_1(X;\hat{\theta})$, $b_1(X;\hat{\gamma})$, $h_1(\tilde{X};\hat{\beta})$ are fit from the trial sample $\mathcal{D}_1$ via polynomial ridge regression with 5-fold cross-validation.

We use trial sizes $n_1 \in \{200, 1000\}$ and $l_x^{\text{FOM}_1} \in \{0.5, 0.2\}$, where a smaller value leads $\text{FOM}_1(X,U)$ to change more quickly in response to X (*i.e.*, consist of high order polynomials), thus resulting in a more complex outcome function $g_1(X)$. We provide examples in Appendix B.3.

When learning $f_1(X)$ from the OS, we conceal U and experiment with $(l_u^{\text{PA}}, \alpha_u^{\text{PA}}) \in \{(\infty,0),(0.5,0),(0.5,10)\}$. In the first setting, $P(A=1|S=2,X,U)$ does not depend on U, and there is *no* hidden confounding, which is intro-

duced when l_u^{PA} is changed to 0.5. Finally setting α_u^{PA} to 10 increases the "weight" of U in $P(A=1|S=2,X,U)$, leading to a larger confounding bias.

The preceding sets of hyperparameters lead to $2\times2\times3=12$ combinations. For each combination, 100 different data-generating functions were sampled from the GPs, leading to 1200 distinct scenarios. For each scenario, 100 independent runs were made where a new trial sample $\mathcal{D}_1$ was generated, and estimates for μ_1 were calculated. An average RMSE is calculated for each scenario over 100 runs and for each combination over 100 scenarios, presented in Figure 4.

5.2. Discussion of Results

We first discuss the results in Figure 1 and focus on the advantages of our prediction-powered estimators over the baselines. We then investigate Figure 2 which demonstrates why the augmented outcome modeling (AOM) is more robust than the additive bias correction (ABC) approach.

Using the OS Alone As the hidden confounding in the OS increases, $\hat{\mu}_1^{\text{OS-OM}}$ in (9), which directly applies $f_1(X)$ in the target population, suffers higher RMSE. Note that its performance does not improve with a larger trial size, as confounding bias does not result from a small sample size.

Using the Trial Sample Alone The performance of $\hat{\mu}_1^{\text{OM}}$ in (4) relies on fitting the outcome function $g_1(X)$ from the trial sample accurately. Therefore, it incurs higher RMSE when the trial is small and $g_1(X)$ is complex. The RMSE gets even worse when higher-order polynomials are used to fit $\hat{g}_1(X)$ in a small trial, due to the quickly increasing variance term (7) (see Appendix B.1).

Combining Trial and Observational Data Prediction-powered estimators $\hat{\mu}_1^{\text{ABC}}$ and $\hat{\mu}_1^{\text{AOM}}$ yield significant improvement over $\hat{\mu}_1^{\text{OM}}$ when the trial is small, outcome function is complex, and the hidden confounding is small. This is because when $f_1(X)$ accurately *estimates away* most of the complex structure in $g_1(X)$, the resulting bias function has a small norm, *i.e.*, $b_1(X)\approx 0$ (see Appendix B.3 for some examples), and is more feasible to fit from the small trial sample and generalize to the target population.

Confounding in the OS leads to slightly worse RMSEs for $\hat{\mu}_1^{\text{ABC}}$ and $\hat{\mu}_1^{\text{AOM}}$, but they still compare favorably to $\hat{\mu}_1^{\text{OM}}$. Note that a larger hidden confounding need not impede benefiting from $f_1(X)$, so long as $b_1(X)$ consists of lower degree polynomials. The results in Figure 4 are *averaged* over many scenarios, and we provide examples in Appendix B.3 where $b_1(X)$ is also "complex" and no improvement over the baseline is achieved. Finally, when the trial is large enough to support fitting $g_1(X)$ with a higher order polynomial, $\hat{\mu}_1^{\text{OM}}$ perform on par with $\hat{\mu}_1^{\text{ABC}}$ and $\hat{\mu}_1^{\text{AOM}}$.

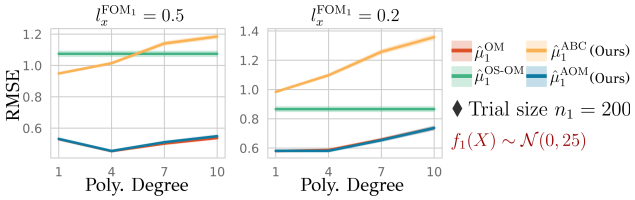

Figure 5. Convention same as Figure 4. The observational predictor is not trained on any data but generates i.i.d. noise for all X.

Which Prediction-powered Estimator is Better? For the experiments in Figure 4, $\hat{\mu}_1^{\text{ABC}}$ and $\hat{\mu}_1^{\text{AOM}}$ have similar performances. We demonstrate the robustness of $\hat{\mu}_1^{\text{AOM}}$ in Figure 5, where $f_1(X)$ is just noise and $b_1(X)$ is harder to estimate than $g_1(X)$. While $\hat{\mu}_1^{\text{ABC}}$ suffers high RMSE, $\hat{\mu}_1^{\text{AOM}}$ simply ignores $f_1(X)$ as a regressor and retains the performance of the baseline approach.

6. Concluding Remarks

We investigated the statistical challenges of generalizing causal inferences from a randomized controlled trial to a target population whose characteristics differ from the trial. We showed how observational data could make generalization more statistically feasible without unrealistic assumptions. Through a diverse set of synthetic experiments, we verified the effectiveness of our methods. Future work includes investigating more flexible approaches to leverage observational data and exploring further experiment setups (*e.g.*, with different kernels to simulate specific real-world settings) to gain further insight into the potentials and limitations of integrating experimental and observational evidence.

Acknowledgements

The authors thank the anonymous reviewers for their thoughtful comments and contributions to our experimental results during the discussion phase, and to Sontag Lab members for insightful discussions. ID was supported by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. This study was supported in part by Office of Naval Research Award No. N00014-21-1-2807.

Impact Statement

Causal inference is crucial in medicine. The present manuscript contributes to the rapidly growing field of integrating gold-standard data from trials and real-world evidence. We propose statistically valid methods that can leverage large-scale observational data to power better generalization of causal effects from a trial to a target population. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023. doi: 10.1126/science.adi6000.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Chen, S., Zhang, B., and Ye, T. Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint (2109.10522)*, September 2021.

Cheng, D. and Cai, T. Adaptive combination of randomized and observational data. *arXiv preprint (2111.15012)*, November 2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1): 165–191, 2024.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.

De Bartolomeis, P., Abad, J., Donhauser, K., and Yang, F. Detecting critical treatment effect bias in small subgroups. *arXiv preprint arXiv:2404.18905*, 2024a.

De Bartolomeis, P., Martinez, J. A., Donhauser, K., and Yang, F. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pp. 1045–1053. PMLR, 2024b.

Degtiar, I., Layton, T., Wallace, J., and Rose, S. Conditional cross-design synthesis estimators for generalizability in medicaid. *Biometrics*, 2023.

Demirel, I., De Brouwer, E., Hussain, Z. M., Oberst, M., Philippakis, A. A., and Sontag, D. Benchmarking observational studies with experimental data under right-censoring. In *International Conference on Artificial Intelligence and Statistics*, pp. 4285–4293, 2024.

Forbes, S. P. and Dahabreh, I. J. Benchmarking observational analyses against randomized trials: a review of studies assessing propensity score methods. *Journal of general internal medicine*, 35:1396–1404, 2020.

Guo, W., Wang, S., Ding, P., Wang, Y., and Jordan, M. I. Multi-source causal inference using control variates. *arXiv preprint arXiv:2103.16689*, 2021.

Han, L., Shen, Z., and Zubizarreta, J. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36: 70453–70482, 2023.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A,*, 178(3):757–778, June 2015.

Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S., and van der Schaar, M. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

Hernan, M. A. and Robins, J. M. *Causal Inference*. CRC Press, Boca Raton, FL, February 2021.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. A structural approach to selection bias. *Epidemiology*, pp. 615–625, 2004.

Hussain, Z., Oberst, M., Shih, M.-C., and Sontag, D. Falsification before extrapolation in causal effect estimation. *arxiv preprint arXiv:2209.13708*, 2022.

Hussain, Z., Shih, M.-C., Oberst, M., Demirel, I., and Sontag, D. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 5869–5898, 2023.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Kallus, N., Puli, A. M., and Shalit, U. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

Karlsson, R. and Krijthe, J. Detecting hidden confounding in observational data using multiple environments.

Advances in Neural Information Processing Systems, 36, 2024.

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

Liao, L. D., Højbjerre-Frandsen, E., Hubbard, A. E., and Schuler, A. Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *arXiv preprint arXiv:2305.19180*, 2023.

NICE. Nice real-world evidence framework, 2022. URL https://www.nice.org.uk/corporate/ecd9/chapter/overview.

Oberst, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D., and Yadlowsky, S. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467*, 2023.

Rasmussen, C. E., Williams, C. K., et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Rosenman, E. T. and Owen, A. B. Designing experiments informed by observational studies. *Journal of Causal Inference*, 9(1):147–171, 2021.

Rosenman, E. T., Basse, G., Owen, A. B., and Baiocchi, M. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 2020.

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

Schuler, A., Walsh, D., Hall, D., Walsh, J., Fisher, C., for Alzheimer's Disease, C. P., Initiative, A. D. N., and Study, A. D. C. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2021.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174 (2):369–386, 2011.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yang, S. and Ding, P. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2019.

Yang, S., Gao, C., Zeng, D., and Wang, X. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85 (3):575–596, 2023.

A. Proofs and Additional Results

A.1. Deferred Proofs

Proposition 3.1. *Let X be a categorical covariate stratifying the population into K groups and denote by $n_{s=1,a,k}$ the number of trial participants from group $X = k$ assigned to treatment $A = a$, and by $\sigma^2_{a,k}$ the variance of outcome among such patients. Let us estimate the outcome function $g_a(X = k)$ with the sample mean of outcomes Y of participants in group $X = k$ assigned to treatment $A = a$.*

Suppose that Assumptions 2.1-2.5 hold. When n_0 is large, the MSE of $\hat{\mu}_a^{\text{OM}}$ in (4) can be approximated as

$$\mathbf{E}[(\hat{\mu}_a^{\text{OM}} - \mu_a)^2] \approx \sum_{k=1}^{K} p^2_{s=0}(k) \frac{\sigma^2_{a,k}}{n_{s=1,a,k}}, \tag{5}$$

where $p_{s=0}(k) \coloneqq P(X = k \mid S = 0)$ is the proportion of patients from group $X = k$ in the target population.

Proof. Let us denote by $\mathcal{D}_{s=1,a,k} \subseteq \mathcal{D}_1$ the set of trial participants in group k that received treatment $A = a$, with size $n_{s=1,a,k} = \sum_{i=1}^{n} \mathbf{1}\{X_i = k, S_i = 1, A_i = a\}$. We estimate the outcome model as

$$\hat{g}_a(X = k) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_i = k, S_i = 1, A_i = a\} Y_i}{n_{s=1,a,k}}. \tag{28}$$

$\hat{g}_a(X = k)$ is simply the sample mean of outcomes Y_i in $\mathcal{D}_{s=1,a,k}$ and we have

$$\begin{aligned}
\mathbf{E}[\hat{g}_a(X = k)] &= \mathbf{E}[Y \mid X = k, S = 1, A = a] \\
&= \mathbf{E}[Y^a \mid X = k, S = 1, A = a] \\
&= \mathbf{E}[Y^a \mid X = k, S = 1] \\
&= \mathbf{E}[Y^a \mid X = k, S = 0]
\end{aligned} \tag{29}$$

where the first equality follows from the unbiasedness of the sample mean and the rest from Assumptions 2.1, 2.2, and 2.4, respectively. When X is categorical and $\hat{g}_a(X)$ is estimated via (28), (4) admits the following equivalent expression.

$$\hat{\mu}_a^{\text{OM}} = \sum_{k=1}^{K} \hat{p}_{s=0}(k) \hat{g}_a(k). \tag{30}$$

where $\hat{p}_{s=0}(k) = \frac{\sum_{i=1}^{n} \mathbf{1}\{S_i=0, X_i=k\}}{\sum_{i=1}^{n} \mathbf{1}\{S_i=0\}}$ is the proportion of patients in the target sample $\mathcal{D}_0$ from group k.

Note that the target $\mathcal{D}_0$ and trial $\mathcal{D}_1$ samples are disjoint of each other. Since $\hat{p}_{s=0}(k)$ is effectively calculated from the observations in the target sample $\mathcal{D}_0$ only, and similarly $\hat{g}_a(k)$ from $\mathcal{D}_1$ only, $\hat{p}_{s=0}(k)$ and $\hat{g}_a(k)$ are independent. Following (30), we can then write

$$\begin{aligned}
\mathbf{E}[\hat{\mu}_a^{\text{OM}}] &= \mathbf{E}\left[\sum_{k=1}^{K} \hat{p}_{s=0}(k) \hat{g}_a(k)\right] \\
&= \sum_{k=1}^{K} \mathbf{E}[\hat{p}_{s=0}(k) \hat{g}_a(k)] \\
&= \sum_{k=1}^{K} \mathbf{E}[\hat{p}_{s=0}(k)] \, \mathbf{E}[\hat{g}_a(k)] \\
&= \sum_{k=1}^{K} p_{s=0}(k) \mathbf{E}[Y^a \mid X = k, S = 0] && (31) \\
&= \mathbf{E}[Y^a \mid S = 0] && \text{(law of total expectation)} \\
&= \mu_a, && (32)
\end{aligned}$$

where (31) follows from the unbiasedness of sample proportion and (29).

Next, by the law of total variance we write

$$\text{Var}\left(\hat{\mu}_a^{\text{OM}}\right) = \mathbf{E}_{\mathcal{D}_0}\left[\text{Var}_{\mathcal{D}_1}(\hat{\mu}_a^{\text{OM}} \mid \mathcal{D}_0)\right] + \text{Var}_{\mathcal{D}_0}\left(\mathbf{E}_{\mathcal{D}_1}\left[\hat{\mu}_a^{\text{OM}} \mid \mathcal{D}_0\right]\right). \tag{33}$$

We start with the first term.

$$\begin{aligned}
\mathbf{E}_{\mathcal{D}_0}\left[\text{Var}_{\mathcal{D}_1}(\hat{\mu}_a^{\text{OM}} \mid \mathcal{D}_0)\right] &= \mathbf{E}_{\mathcal{D}_0}\left[\text{Var}_{\mathcal{D}_1}\left(\sum_{k=1}^{K}\hat{p}_{s=0}(k)\hat{g}_a(k)\Big|\mathcal{D}_0\right)\right] \\
&= \mathbf{E}_{\mathcal{D}_0}\left[\sum_{k=1}^{K}\hat{p}^2_{s=0}(k)\text{Var}_{\mathcal{D}_1}\left(\hat{g}_a(k)\big|\mathcal{D}_0\right)\right] &(34)\\
&= \sum_{k=1}^{K}\mathbf{E}_{\mathcal{D}_0}\left[\hat{p}^2_{s=0}(k)\text{Var}_{\mathcal{D}_1}\left(\hat{g}_a(k)\big|\mathcal{D}_0\right)\right] \\
&= \sum_{k=1}^{K}\mathbf{E}_{\mathcal{D}_0}\left[\hat{p}^2_{s=0}(k)\right]\text{Var}_{\mathcal{D}_1}\left(\hat{g}_a(k)\right) &(35)\\
&= \sum_{k=1}^{K}\left(\mathbf{E}_{\mathcal{D}_0}\left[\hat{p}_{s=0}(k)\right]^2 + \underbrace{\text{Var}_{\mathcal{D}_0}\left(\hat{p}_{s=0}(k)\right)}_{\xrightarrow{n_0\to\infty}0}\right)\text{Var}_{\mathcal{D}_1}\left(\hat{g}_a(k)\right) &(36)\\
&\approx \sum_{k=1}^{K}p^2_{s=0}(k)\frac{\sigma^2_{a,k}}{n_{s=1,a,k}}, &(37)
\end{aligned}$$

where (34) holds since the participants in different groups are independent of each other and $\hat{p}^2_{s=0}(k)$ is no longer random after conditioning on $\mathcal{D}_0$. Similar to above, (35) holds since $\hat{g}_a(k)$ is independent of the target sample $\mathcal{D}_0$ and therefore $\hat{p}^2_{s=0}(k)$. (36) follows after writing the variance of the sample proportion

$$\text{Var}_{\mathcal{D}_0}\left(\hat{p}_{s=0}(k)\right) = \frac{p_{s=0}(k)\left(1-p_{s=0}(k)\right)}{n_0}. \tag{38}$$

For the second term we write

$$\begin{aligned}
\text{Var}_{\mathcal{D}_0}\left(\mathbf{E}_{\mathcal{D}_1}\left[\hat{\mu}_a^{\text{OM}} \mid \mathcal{D}_0\right]\right) &= \text{Var}_{\mathcal{D}_0}\left(\mathbf{E}_{\mathcal{D}_1}\left[\sum_{k=1}^{K}\hat{p}_{s=0}(k)\hat{g}_a(k) \mid \mathcal{D}_0\right]\right) \\
&= \text{Var}_{\mathcal{D}_0}\left(\sum_{k=1}^{K}\hat{p}_{s=0}(k)\mathbf{E}_{\mathcal{D}_1}\left[\hat{g}_a(k) \mid \mathcal{D}_0\right]\right) \\
&= \text{Var}_{\mathcal{D}_0}\left(\sum_{k=1}^{K}\hat{p}_{s=0}(k)\mathbf{E}_{\mathcal{D}_1}\left[\hat{g}_a(k)\right]\right) \\
&= \text{Var}_{\mathcal{D}_0}\left(\sum_{k=1}^{K}\hat{p}_{s=0}(k)\mathbf{E}\left[Y^a \mid X=k, S=0\right]\right) \\
&= \sum_{k=1}^{K}\underbrace{\text{Var}_{\mathcal{D}_0}\left(\hat{p}_{s=0}(k)\right)}_{\xrightarrow{n_0\to\infty}0}\mathbf{E}\left[Y^a \mid X=k, S=0\right]^2 &(39)\\
&\approx 0, &(40)
\end{aligned}$$

where (39) follows again from (38). Combining (33), (37), and (40), we have, as n_0 goes to infinity,

$$\text{Var}\left(\hat{\mu}_a^{\text{OM}}\right) \approx \sum_{k=1}^{K}p^2_{s=0}(k)\frac{\sigma^2_{a,k}}{n_{s=1,a,k}} \tag{41}$$

Finally, we have

$$\begin{aligned}\mathbf{E}[(\hat{\mu}_a^{\mathrm{OM}} - \mu_a)^2] &= \underbrace{\mathbf{E}[(\hat{\mu}_a^{\mathrm{OM}} - \mu_a)]^2}_{0 \text{ by } (32)} + \mathrm{Var}\left(\hat{\mu}_a^{\mathrm{OM}}\right) \\ &\approx \sum_{k=1}^{K} p_{s=0}^2(k) \frac{\sigma_{a,k}^2}{n_{s=1,a,k}}, \qquad \text{(by (41))}\end{aligned}$$

and we are done. □

Theorem 3.2. *Suppose that Assumptions 2.1-2.5 hold and consider a parametric estimator $\hat{g}_a(X) = g_a(X;\hat{\theta})$ for the outcome function. For large n_0, the MSE of $\hat{\mu}_a^{\mathrm{OM}}$ in (4) can be approximated as*

$$\begin{aligned}&\mathbf{E}[(\hat{\mu}_a^{\mathrm{OM}} - \mu_a)^2] \\ &\approx \mathbf{E}_{X\sim P_0}\big[\underbrace{\mathbf{E}_{\hat{\theta}\sim\mathcal{A}(P_1)}[g_a(X;\hat{\theta})] - g_a(X)}_{=:\ \mathrm{SB}_g(X)}\big]^2 \qquad (6) \\ &\quad + \mathrm{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\big(\mathbf{E}_{X\sim P_0}[g_a(X;\hat{\theta})]\big). \qquad (7)\end{aligned}$$

Proof.

$$\mathbf{E}[\big(\hat{\mu}_a^{\mathrm{OM}} - \mu_a\big)^2] = \underbrace{\big(\mathbf{E}\left[\hat{\mu}_a^{\mathrm{OM}}\right] - \mu_a\big)^2}_{\mathrm{Bias}^2} + \underbrace{\mathrm{Var}(\hat{\mu}_a^{\mathrm{OM}})}_{\mathrm{Variance}}. \qquad (42)$$

We will start with the bias term. Note that, under Assumptions 2.1-2.5, we have

$$\begin{aligned}\mu_a &= \mathbf{E}[\mathbf{E}[Y \mid X, S=1, A=a] \mid S=0] && \text{(see (2))} \\ &= \mathbf{E}[g_a(X) \mid S=0] && \text{(by definition, see (3))} \\ &= \mathbf{E}_{X\sim P_0}[g_a(X)]. && (43)\end{aligned}$$

with a manipulation of notation at the final step. Once $\hat{\theta}$ is estimated from the trial sample $\mathcal{D}_1$ via an algorithm $\mathcal{A}$, $\hat{\mu}_a^{\mathrm{OM}}$ is calculated by effectively taking a sample mean of $g_a(X_i;\hat{\theta})$ for the covariates X_i in the target sample $\mathcal{D}_0$. We can write,

$$\begin{aligned}\mathbf{E}[\hat{\mu}_a^{\mathrm{OM}}] &= \mathbf{E}\left[\frac{1}{n_0}\sum_{i=1}^{n} \mathbf{1}\left\{S_i = 0\right\} g_a(X_i;\hat{\theta})\right] \\ &= \mathbf{E}\left[\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\right] \\ &= \mathbf{E}_{\hat{\theta}\sim P_1}\left[\mathbf{E}_{\mathcal{D}_0}\left[\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\Big|\hat{\theta}\right]\right] \\ &= \mathbf{E}_{\hat{\theta}\sim P_1}\left[\mathbf{E}_{\mathcal{D}_0}\left[\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\right]\right] \\ &= \mathbf{E}_{\hat{\theta}\sim P_1}\left[\mathbf{E}_{X\sim P_0}\left[g_a(X;\hat{\theta})\right]\right] \\ &= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}_{\hat{\theta}\sim P_1}\left[g_a(X;\hat{\theta})\right]\right]. \qquad (44)\end{aligned}$$

since $X_i \in \mathcal{D}_0$ are i.i.d and independent of $\hat{\theta}$. Combining (43) and (44) we write

$$\begin{aligned}\mathrm{Bias} &= \mathbf{E}\left[\hat{\mu}_a^{\mathrm{OM}}\right] - \mu_a \\ &= \mathbf{E}_{X\sim P_0}\Big[\mathbf{E}_{\hat{\theta}\sim\mathcal{A}(P_1)}[g_a(X;\hat{\theta})] - g_a(X)\Big]. \qquad (45)\end{aligned}$$

We continue with the variance term by invoking the law of total variance.

$$\begin{aligned}
\text{Variance} &= \text{Var}\left(\hat{\mu}_a^{\text{OM}}\right) \\
&= \text{Var}_{\hat{\theta}\sim\mathcal{A}(P_1),\mathcal{D}_0}\left(\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\right) \\
&= \text{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{\mathcal{D}_0}\left[\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\middle|\hat{\theta}\right]\right) \\
&\qquad + \underbrace{\mathbf{E}_{\hat{\theta}\sim\mathcal{A}(P_1)}\left[\underbrace{\text{Var}_{\mathcal{D}_0}\left(\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\middle|\hat{\theta}\right)}_{\xrightarrow{n_0\to\infty}0}\right]}_{\approx 0} \\
&\approx \text{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{\mathcal{D}_0}\left[\frac{1}{n_0}\sum_{X_i\in\mathcal{D}_0} g_a(X_i;\hat{\theta})\right]\right) \\
&= \text{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}\left[g_a(X;\hat{\theta})\right]\right), \qquad (46)
\end{aligned}$$

where in the third equality, the variance of the sample mean vanishes for large n_0. Last two steps follow since the target sample $\mathcal{D}_0$ and $\hat{\theta}$ are independent and $X_i \in \mathcal{D}_0$ are i.i.d. Plugging (45) and (46) into the definition of the MSE in (42), we are done. □

Lemma 4.1. *Suppose that Assumptions 2.1-2.5 hold. Let $f_a : \mathcal{X} \to \mathbb{R}$ and define the error variable*

$$Z \coloneqq f_a(X) - Y. \qquad (12)$$

μ^a can be identified as

$$\mu_a = \mathbf{E}_{X\sim P_0}[f_a(X) - \mathbf{E}[Z \mid X, S=1, A=a]]. \qquad (13)$$

Proof. Recall that we have

$$\mu_a = \mathbf{E}\left[f_a(X) \mid S=0\right] - \mathbf{E}\left[f_a(X) - Y^a \mid S=0\right]. \qquad (47)$$

Note that the prediction model f_a is fixed. Therefore we have

$$\mathbf{E}[f_a(X) \mid S=0] = \mathbf{E}_{X\sim P_0}[f_a(X)], \qquad (48)$$

which is only a change of notation. Further, conditioned on X, $f_a(X)$ is no more random, We can then write

$$\begin{aligned}
\mathbf{E}[f_a(X) - Y^a \mid S=0] &= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}[f_a(X) - Y^a \mid X, S=0]\right] \\
&= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}[f_a(X) - Y^a \mid X, S=1]\right] \qquad (49) \\
&= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}[f_a(X) - Y^a \mid X, S=1, A=a]\right] \qquad (50) \\
&= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}[f_a(X) - Y \mid X, S=1, A=a]\right] \qquad (51) \\
&= \mathbf{E}_{X\sim P_0}\left[\mathbf{E}[Z \mid X, S=1, A=a]\right], \qquad (52)
\end{aligned}$$

where (49) is due to Assumptions 2.4 and 2.5, (50) is due to Assumptions 2.2 and 2.3, and (51) is due to Assumption 2.1.

Combining (47), (48), and (52) completes the proof. □

Theorem 4.2. *Suppose that Assumptions 2.1-2.5 hold. For large n_0, the MSE of $\hat{\mu}_a^{\text{ABC}}$ in (14) can be approximated as*

$$\begin{aligned}
&\mathbf{E}[(\hat{\mu}_a^{\text{ABC}} - \mu_a)^2] \\
&\approx \mathbf{E}_{X\sim P_0}\left[\mathbf{E}_{\hat{\gamma}\sim\mathcal{A}(P_1)}[b_a(X;\hat{\gamma})] - b_a(X)\right]^2 \qquad (15) \\
&\quad + \text{Var}_{\hat{\gamma}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}[b_a(X;\hat{\gamma})]\right). \qquad (16)
\end{aligned}$$

Proof. We have, by Lemma 4.1,

$$\begin{aligned}\mu_a &= \mathbf{E}[f_a(X) - b_a(X) \mid S = 0] \\ &= \mathbf{E}[f_a(X) - b_a(X) \mid S = 0].\end{aligned} \tag{53}$$

where

$$b_a(X) = \mathbf{E}[f_a(X) - Y \mid X, S = 1, A = a].$$

We consider a parametric estimator $b_a(X; \hat{\gamma})$ where γ is estimated from the instance-wise prediction errors $f_a(X) - Y$ in the trial sample $\mathcal{D}_1$.

We will follow the same steps in the proof of Theorem 3.2 for the most part.

$$\mathbf{E}[(\hat{\mu}_a^{\text{ABC}} - \mu_a)^2] = \underbrace{\left(\mathbf{E}\left[\hat{\mu}_a^{\text{ABC}}\right] - \mu_a\right)^2}_{\text{Bias}^2} + \underbrace{\text{Var}(\hat{\mu}_a^{\text{ABC}})}_{\text{Variance}}. \tag{54}$$

We have

$$\begin{aligned}\mathbf{E}[\hat{\mu}_a^{\text{ABC}}] &= \mathbf{E}\left[\frac{1}{n_0}\sum_{i=1}^{n} \mathbf{1}\left\{S_i = 0\right\}\left(f_a(X_i) - b^a(X_i; \hat{\gamma})\right)\right] \\ &= \mathbf{E}\left[\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} f_a(X_i) - b^a(X_i; \hat{\gamma})\right].\end{aligned} \tag{55}$$

Since the sample mean is unbiased, it follows that

$$\mathbf{E}\left[\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} f_a(X_i)\right] = \mathbf{E}_{X \sim P_0}\left[f_a(X)\right]. \tag{56}$$

Next, via the same machinery that derives (44), we have

$$\mathbf{E}\left[\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} b^a(X_i, \hat{\gamma})\right] = \mathbf{E}_{X \sim P_0}\Big[\mathbf{E}_{\hat{\gamma} \sim P_1}\left[b_a(X; \hat{\gamma})\right]\Big]. \tag{57}$$

By (55), (56), and (57), we observe

$$\mathbf{E}[\hat{\mu}_a^{\text{ABC}}] = \mathbf{E}_{X \sim P_0}\Big[f_a(X) - \mathbf{E}_{\hat{\gamma} \sim P_1}\left[b_a(X; \hat{\gamma})\right]\Big], \tag{58}$$

which leads to, in combination with (53)

$$\begin{aligned}\text{Bias} &= \mathbf{E}\left[\hat{\mu}_a^{\text{ABC}}\right] - \mu_a \\ &= \mathbf{E}_{X \sim P_0}\Big[b_a(X) - \mathbf{E}_{\hat{\gamma} \sim P_1}\left[b_a(X; \hat{\gamma})\right]\Big].\end{aligned} \tag{59}$$

We continue with the variance term.

$$\begin{aligned}\text{Variance} &= \text{Var}\left(\hat{\mu}_a^{\text{ABC}}\right) \\ &= \text{Var}\left(\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} f_a(X_i) - b^a(X_i; \hat{\gamma})\right) \\ &= \underbrace{\text{Var}_{\mathcal{D}_0}\left(\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} f_a(X_i)\right)}_{\xrightarrow{n_0 \to \infty} 0} \\ &\quad + \text{Var}_{\hat{\gamma} \sim \mathcal{A}(P_1), \mathcal{D}_0}\left(\frac{1}{n_0}\sum_{X_i \in \mathcal{D}_0} b^a(X_i; \hat{\gamma})\right)\end{aligned}$$

$$\approx \text{Var}_{\hat{\gamma}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}\left[b_a(X;\hat{\gamma})\right]\right). \tag{60}$$

The decomposition in third equality is due to the independence of the models predictions $f_a(X)$ and errors $f_a(X) - Y$ ($\hat{\gamma}$ is derived using the latter only). Finally, (60) follows through the same machinery that derives (46).

We are done after plugging (59) and (60) into (54). □

A.2. Polynomial Ridge Regression

We consider polynomial ridge regression in the trial sample $\mathcal{D}_1$ using Legendre polynomials up to degree d', to fit the outcome model and bias model estimates, $\hat{g}_a$ and $\hat{b}_a$ which results in the following fits with an appropriately chosen penalty parameter λ (Wainwright, 2019).

$$\hat{g}_a \in \underset{g\in\mathcal{F}(d')}{\text{argmin}} \left\{ \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - g(X_i))^2 \right\}, \tag{61}$$

$$\hat{b}_a \in \underset{b\in\mathcal{F}(d')}{\text{argmin}} \left\{ \frac{1}{n_1}\sum_{i=1}^{n_1}(Z_i - b(X_i))^2 \right\}. \tag{62}$$

where

$$\mathcal{F}(d') \coloneqq \left\{ \sum_{k=0}\beta_k\phi_k(X) \mid \sum_{k=0}^{d'}\beta_k^2 \leq 1 \right\}, \tag{63}$$

is the class of polynomials up to degree d' with bounded norm. The results then follow from the oracle inequalities derived for the orthogonal basis approximation problem in Example 13.14, Section 13.3 of (Wainwright, 2019).

A.3. MSE Approximation for the Augmented Outcome Modeling Approach

Theorem A.1. *Suppose that Assumptions 2.1-2.5 hold. For large n_0, the MSE of $\hat{\mu}_a^{\text{AOM}}$ in (25) can be approximated as*

$$\mathbf{E}[(\hat{\mu}_a^{\text{AOM}} - \mu_a)^2] \approx \mathbf{E}_{X\sim P_0}\left[\mathbf{E}_{\hat{\beta}\sim\mathcal{A}(P_1)}[h_a(\tilde{X};\hat{\beta})] - h_a(\tilde{X})\right]^2 + \text{Var}_{\hat{\beta}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}[h_a(\tilde{X};\hat{\beta})]\right). \tag{64}$$

Proof. The proof follows from the same steps in the proof of Theorem 3.2. □

A.4. Doubly-Robust Estimation

In order to leverage the prognostic model $f_a(X)$ in the analysis, we can proceed with two identifications of μ_a, (13) and (24), for which we considered estimators based *only* on regression functions ((14) and (25)). However, in practice, we can directly use the so-called doubly-robust (DR) estimators for (13) and (24), which have several desirable properties.

In addition to a regression function component, DR estimators also have *weighting* function components, which, in our case, are the probability of trial enrollment, $P(S = 1 \mid X)$, and the probability of treatment assignment in the trial, $P(A = a \mid X, S = 1)$. Estimators based only on *weighting* models are also available but will not be covered here in the interest of space. (Dahabreh et al., 2020) derives a generic DR estimator for the functional

$$\mathbf{E}_{X\sim P_0}[\mathbf{E}[Y \mid X, S = 1, A = a]], \tag{65}$$

which we can directly adopt and use to estimate (24) and the second term of (13). Estimating the first term of (13) remains unchanged as the average of predictions $f_a(X)$ in the target sample. We make the following definitions.

$$p = P(S = 1). \tag{66}$$

$$p(X) = P(S = 1 \mid X). \tag{67}$$

$$\pi_a(X) = P(A = a \mid X, S = 1), \tag{68}$$

and denote by $\hat{p}$, $\hat{p}(X)$, and $\hat{\pi}_a(X)$ their estimates. Note that in order to use DR estimators, one now needs to fit those functions using the composite sample $\mathcal{D}$. Next we give the DR estimator for (13) and (24).

$$\hat{\mu}_a^{\text{DR-ABC}} = \frac{1}{n_0}\sum_{i=1}^{n}\mathbf{1}\left\{S_i = 0\right\} f_a(X_i)$$

$$+\frac{1}{n(1-\hat{p})}\sum_{i=1}^{n}\left(\mathbf{1}\left\{S_i=0\right\}\hat{b}_a(X_i)+\mathbf{1}\left\{S_i=1,A=a\right\}\frac{1-\hat{p}(X_i)}{\hat{p}(X_i)\hat{\pi}^a(X_i)}\big(Z_i-\hat{b}_a(X_i)\big)\right). \tag{69}$$

$$\hat{\mu}^a_{\text{DR-PA}}=\frac{1}{n(1-\hat{p})}\sum_{i=1}^{n}\left(\mathbf{1}\left\{S_i=0\right\}\hat{h}_a(\tilde{X}_i)+\mathbf{1}\left\{S_i=1,A=a\right\}\frac{1-\hat{p}(X_i)}{\hat{p}(\tilde{X}_i)\hat{\pi}^a(\tilde{X}_i)}\big(Y-\hat{h}_a(\tilde{X}_i)\big)\right). \tag{70}$$

An essential property of DR estimators is that consistent estimation of *either* the regression or weighting functions guarantees consistent estimation of μ_a, hence the name "doubly-robust". Beyond this DR property, however, they have other desirable properties (under certain regularity conditions or cross-fitting techniques (Chernozhukov et al., 2018)) such as asymptotical efficiency and normality, which enable one to construct confidence intervals beyond point prediction and allow for, *e.g.*, calculating p-values and testing hypotheses. We refer the interested reader to (Kennedy, 2023) for a unifying overview of the theory around the DR estimators, their properties, and how to construct them for different estimands of interest. We present empirical results for the DR estimators in Appendix B.2.

B. Additional Experimental Results

B.1. Bias-Variance Tradeoff

We do not plot the bias and variance terms for $\hat{\mu}_1^{\text{OS-OM}}$. It has minimal ($\approx 0$) variance as one applies the observational predictor directly to the target sample, and nothing is fit from the small trial data. Since the target sample is taken to be large, the variance in $\hat{\mu}_1^{\text{OS-OM}}$ is negligible. Almost all of its MSE (see Figure 4) consists of the bias, which results from hidden confounding introduced by concealing the U variable (see Figure 3).

In Figure 6, we see that the bias resulting from estimating the outcome function $g_1(X)$ from the trial sample is very large with a small model. Although it decreases with a larger model as expected, we see in Figure 7 that the variance quickly explodes when the trial size is small ($n_1 = 200$) and $g_1(X)$ is complex and has high intrinsic variation. When$n_1 = 1000$, the variance terms significantly decrease by a factor of 10, and the RMSE of $\hat{\mu}_1^{\text{OM}}$ becomes comparable to prediction-powered estimators when higher-degree polynomials are fit for $\hat{g}_1(X)$.

Our approaches leveraging the additional predictor have smaller bias and variance terms. The difference is the most significant when the trial is small and the outcome function is complex.

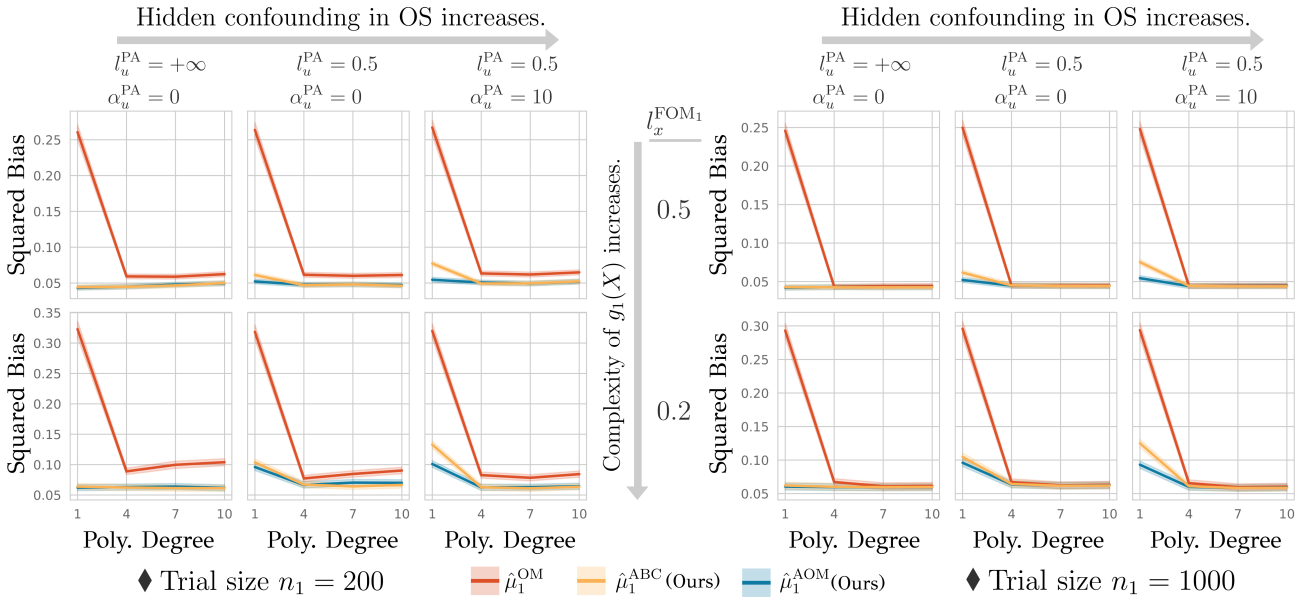

Figure 6. Convention same as Figure 4. Average squared bias of each estimator is plotted.

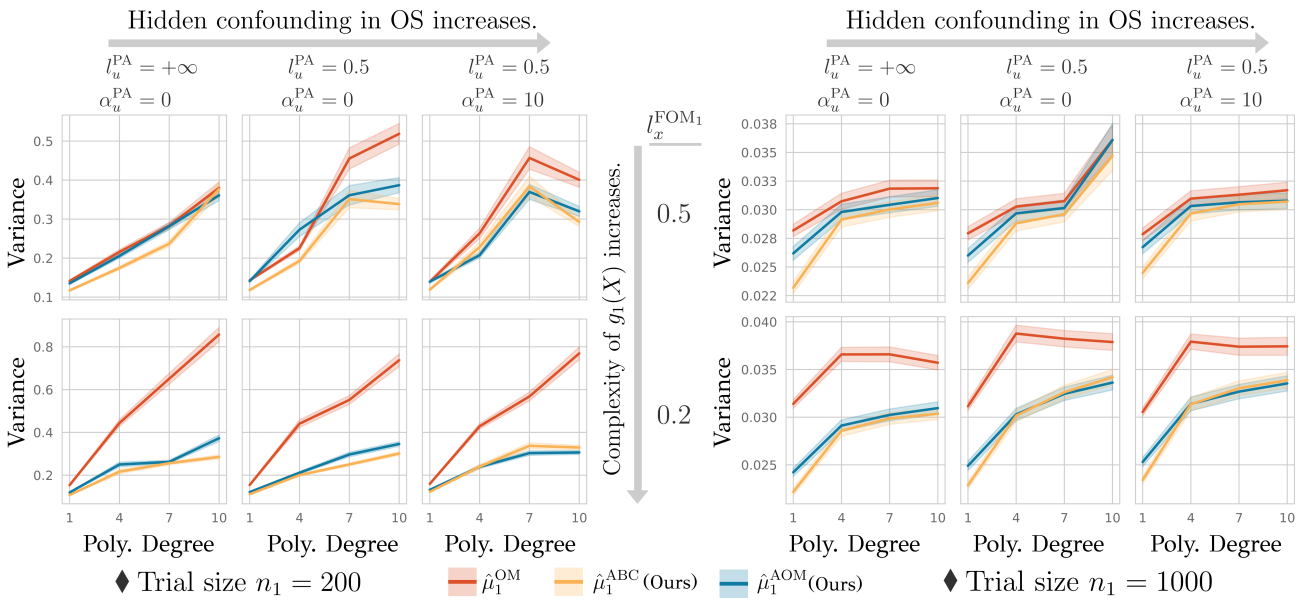

Figure 7. Convention same as Figure 4. Average variance of each estimator is plotted.

B.2. IPW and DR-based Estimators

In Figure 8, we include the generalization RMSE for the doubly-robust (DR) and inverse propensity weighting (IPW) estimators. DR versions of our methods are given in Appendix A.4. Baseline IPW and DR estimators are detailed in (Dahabreh et al., 2020). The IPW estimator performs the worst due to the high variance in the propensity weight estimates, and the DR estimators perform similarly to the outcome-model (OM) estimators. Dahabreh et al. (2020) report similar results. Finally, we note that as the sample size in the trial n_1 increases, the MSE of different estimators converge as before.

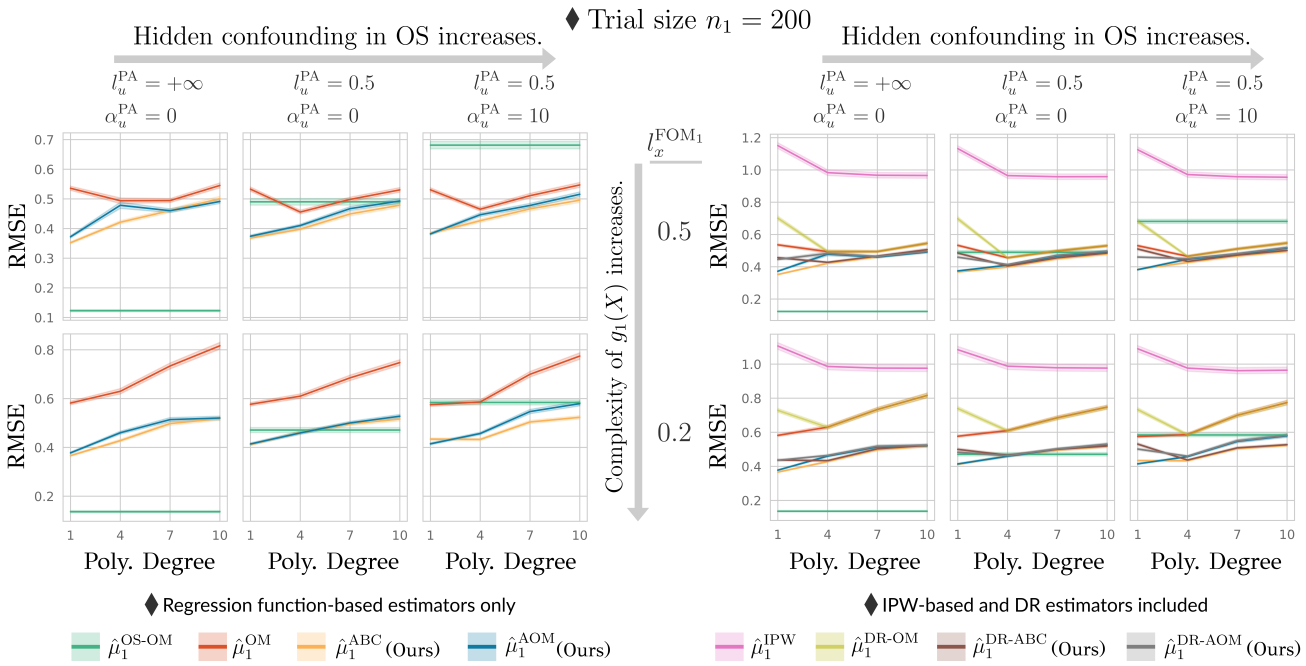

Figure 8. Generalization RMSE with DR and IPW estimators included.

B.3. Example Cases

In Figures 9-12, we demonstrate several example cases where we plot the ground truth functions for the synthetic data-generating processes, an example trial sample that is used to fit $\hat{b}_1(X)$, $\hat{g}_1(X)$, $\hat{h}_1(X)$ (plotted the linear fits only for simplicity, *i.e.*, 1st-degree polynomials), and the observational predictor $f_1(X)$.

We aim to demonstrate how the outcome function $g_1(X)$ becomes "wiggly" as $l_x^{\text{FOM}_1}$ decreases and has more rapid turns representing higher-order polynomials. Further, we see that as the hidden confounding increases, *i.e.*, as we move along the x-axis of plots, the bias of the observational predictor, $b_1(X) = g_1(X) - f_1(X)$ also increases and becomes a "higher-norm" function, decreasing the utility of leveraging observational data and increasing the RMSE.

As we referred to earlier, one can see in Figures 11 and 12, for the cases with $l_x^{\text{FOM}_1} = 0.2$ (complex outcome function $g_1(X)$) and ($l_u^{\text{PA}} = 0.5, \alpha_u^{\text{FOM}_1} = 10$) (large hidden confounding), the bias function $b_1(X)$ is also a complex function with high norm, and the RMSE of the prediction-powered approaches are not significantly better than the baseline estimator $\hat{\mu}_1^{\text{OM}}$.

B.4. Using (Generalized) Linear Models in the Data-generating Process

We sincerely thank Reviewer xwX2 for taking the time during the author-reviewer discussion phase to provide the initial codebase for the results presented here. Instead of generating the outcome and propensity score functions using GPs, we use polynomial models.

The full outcome model under treatment $A = 1$ is specified as a 5-th order polynomial with parameter β. Precisely, we set

$$\text{FOM}_1(X, U) = \beta_0 + \beta_1^X X + \cdots + \beta_5^X X^5 + \gamma \left(\beta_1^U U + \cdots + \beta_5^U U^5 + \beta_1^{XU} XU + \cdots + \beta_5^{XU} (XU)^5\right), \tag{71}$$

and observe $Y^1 = \text{FOM}_1(X, U) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma \in \mathbb{R}_+$ which we use to model the intrinsic variation in outcome observations. Larger values for σ increase the risk of overfitting when the trial size n_1 is small. β parameters characterize the complexity of the outcome function.

The probability of selection into the trial and the probability of treatment assignment in the OS are modeled as follows.

$$P(S = 1|X, U) = \frac{1}{1 + \exp\left(\lambda_0 + \lambda_1 X + \cdots + \lambda_5 X^5\right)}. \tag{72}$$

$$P(A = 1|S = 2, X, U) = \frac{1}{1 + \exp\left(\alpha_0 + \alpha_1^X X + \cdots + \alpha_5^X X^5 + \gamma \left(\alpha_1^U U + \cdots + \alpha_5^U U^5 + \alpha_1^{XU} XU + \cdots + \alpha_5^{XU} (XU)^5\right)\right)}. \tag{73}$$

$\gamma \in \mathbb{R}_+$ determines the amount of hidden confounding in the OS, as U is concealed. Further, λ parameters characterize how weak the overlap is between the trial and target samples.

Briefly, larger values for β, γ, λ, and σ parameters make the generalization task more challenging. In Table 1, we present the generalization MSEs under various settings. We always use $\alpha \sim \mathcal{N}(0, 1)$. We sample 100 ground-truth α, β parameters for each setting, make 100 independent runs for each ground-truth, and then present the average MSE values. We use both 1st and 5th order polynomials to fit the bias and outcome functions, $\hat{b}_1(X)$ and $\hat{g}_1(X)$.

Setting				1st order poly. fit		5th order poly. fit	
				$\hat{\mu}_1^{\text{ABC}}$	$\hat{\mu}_1^{\text{OM}}$	$\hat{\mu}_1^{\text{ABC}}$	$\hat{\mu}_1^{\text{OM}}$
$\gamma = 0$,	$\epsilon \sim \mathcal{N}(0, 0.1^2)$,	$\beta \sim \mathcal{N}(0, 1)$,	$\lambda = 1$	**.0001**	.0092	.0002	.0002
$\gamma = 1$,	$\epsilon \sim \mathcal{N}(0, 0.1^2)$,	$\beta \sim \mathcal{N}(0, 1)$,	$\lambda = 1$	**.0087**	.0183	.0148	.0148
$\gamma = 0$,	$\epsilon \sim \mathcal{N}(0, 2^2)$,	$\beta \sim \mathcal{N}(0, 1)$,	$\lambda = 1$	**.0044**	.0054	.0066	.0066
$\gamma = 0$,	$\epsilon \sim \mathcal{N}(0, 2^2)$,	$\beta \sim \mathcal{N}(0, 2^2)$,	$\lambda = 1$	**.0044**	.0082	.0066	.0066
$\gamma = 0$,	$\epsilon \sim \mathcal{N}(0, 2^2)$,	$\beta \sim \mathcal{N}(0, 2^2)$,	$\lambda = 2$	**.0048**	.0127	.0151	.0151
$\gamma = 1$,	$\epsilon \sim \mathcal{N}(0, 2^2)$,	$\beta \sim \mathcal{N}(0, 2^2)$,	$\lambda = 2$	**.0060**	.0141	.0139	.0139

Table 1. Generalization MSEs using (generalized) linear models in the data-generating process.

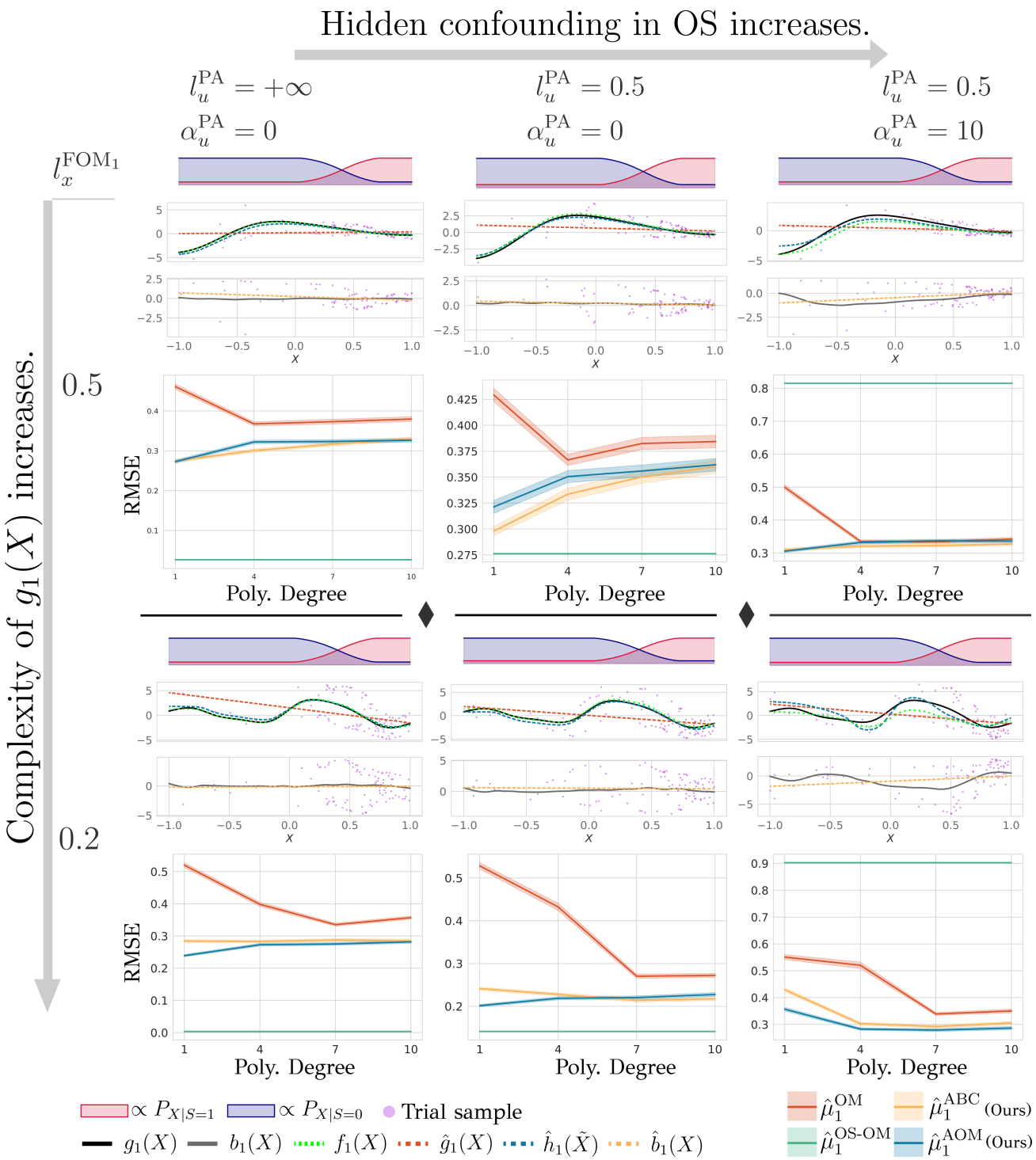

Figure 9. Example case 1.

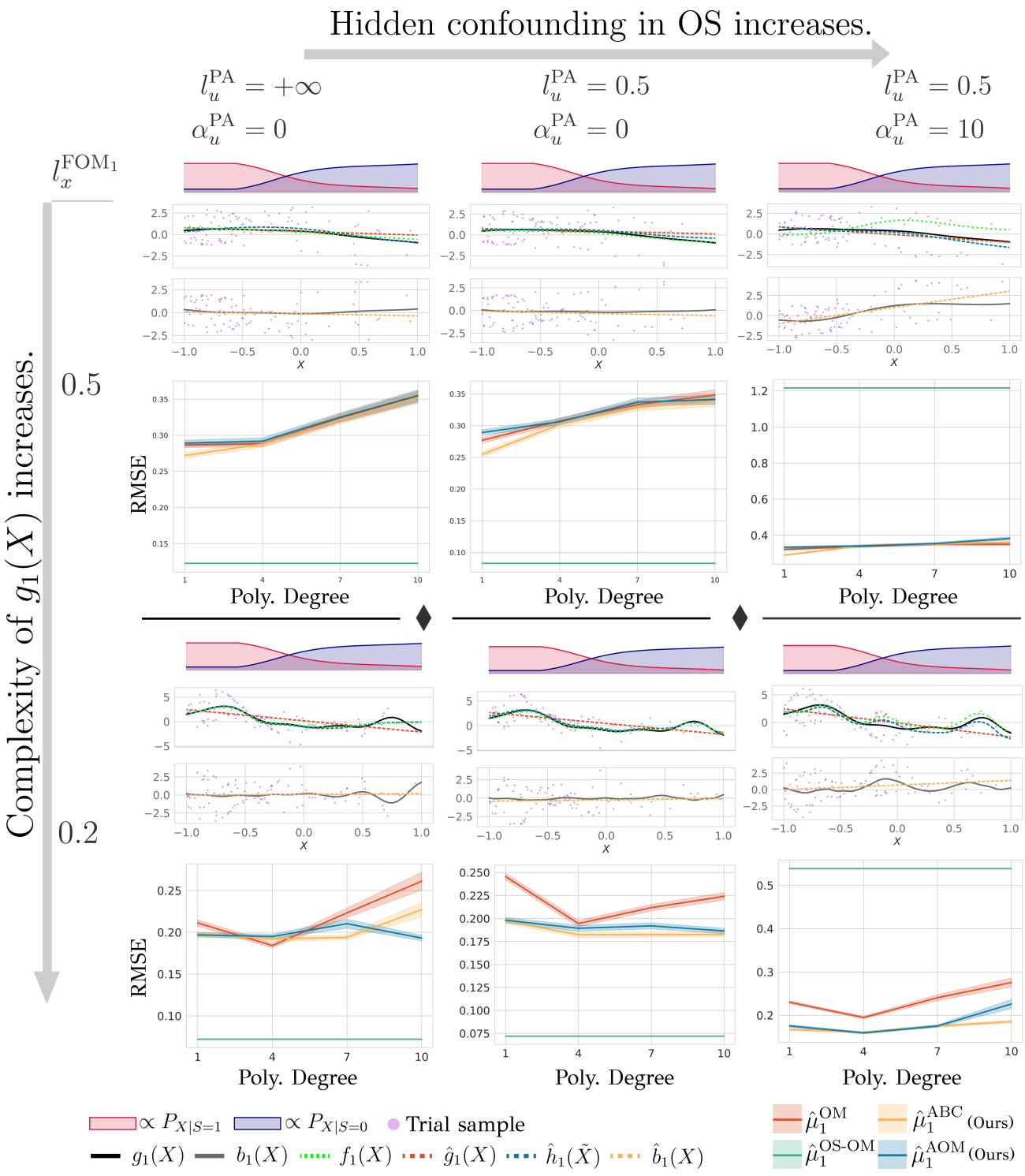

Figure 10. Example case 2.

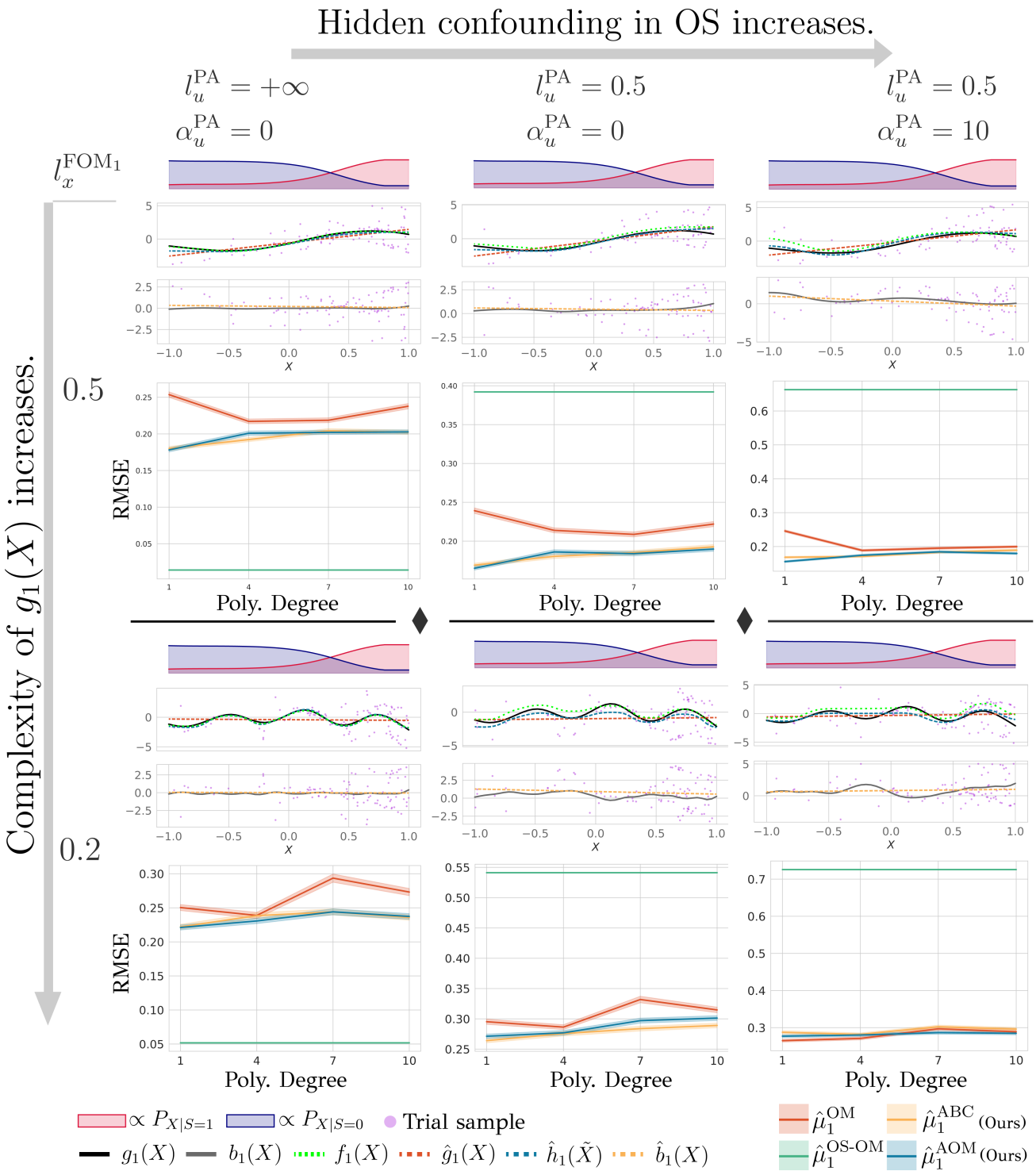

Figure 11. Example case 3.

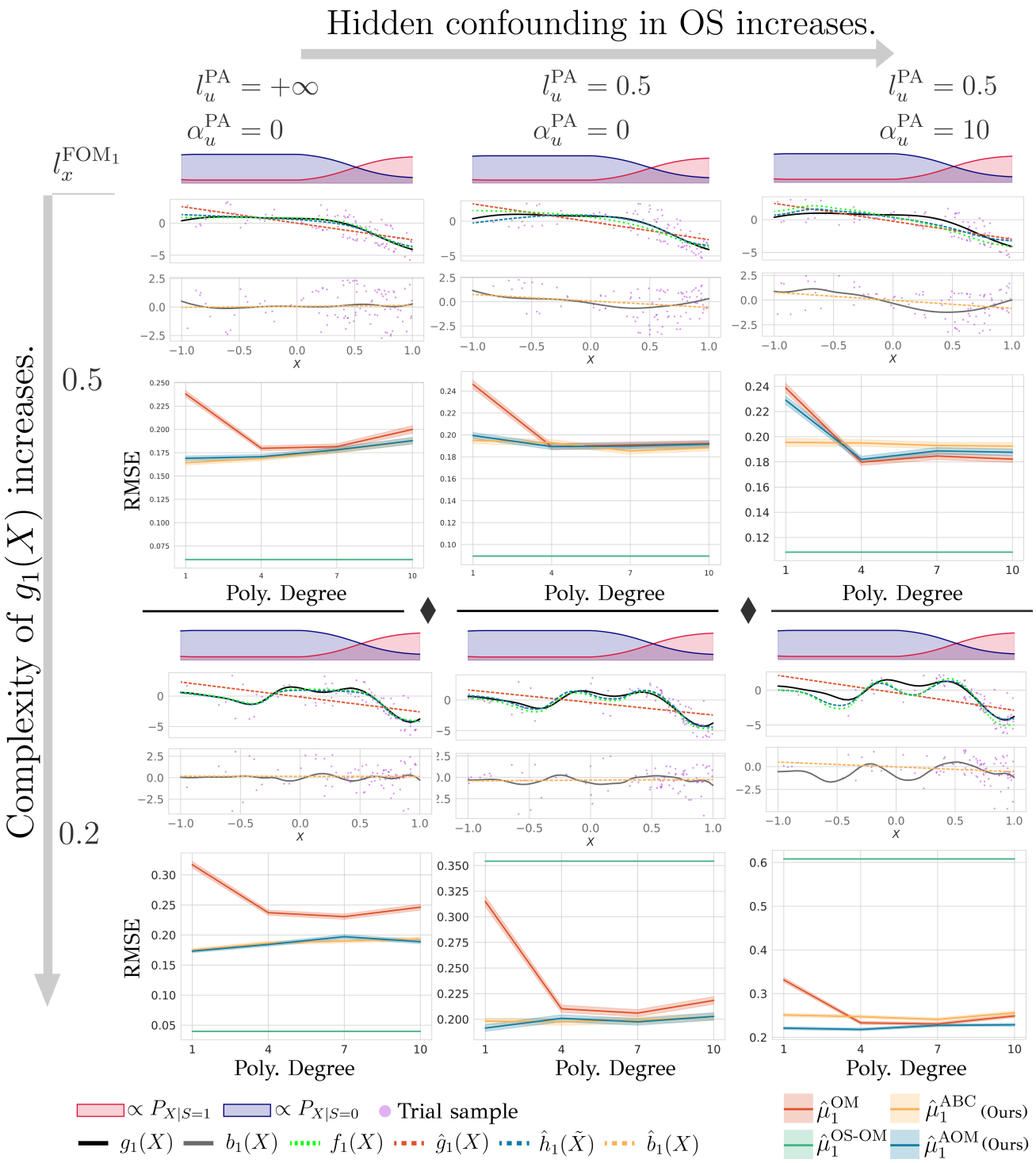

Figure 12. Example case 4.