INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Observational vs. Experimental Data When Making Automated Decisions Using Machine Learning

Carlos Fernández-Loría, Foster Provost

To cite this article:
Carlos Fernández-Loría, Foster Provost (2025) Observational vs. Experimental Data When Making Automated Decisions Using Machine Learning. INFORMS Journal on Data Science 4(3):197-229. https://doi.org/10.1287/ijds.2023.0012

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 4, No. 3, July–September 2025, pp. 197–229
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Observational vs. Experimental Data When Making Automated Decisions Using Machine Learning

Carlos Fernández-Loría,[a,*] **Foster Provost**[b]

[a] School of Business and Management, Hong Kong University of Science and Technology, New Territories, Hong Kong; [b] Leonard N. Stern School of Business, New York University, New York, New York 10012

*Corresponding author

Contact: imcarlos@ust.hk, https://orcid.org/0000-0003-4509-3768 (CF-L); fprovost@stern.nyu.edu (FP)

Received: June 19, 2023
Revised: August 27, 2024; March 9, 2025
Accepted: March 21, 2025
Published Online in Articles in Advance: June 3, 2025

https://doi.org/10.1287/ijds.2023.0012

Abstract. Decisions supported by machine learning often aim to improve outcomes through interventions, such as influencing purchasing behavior with ads or increasing customer retention with special offers. However, using observational data to estimate these effects can introduce confounding bias. Although experimental data can mitigate confounding, it is not always feasible to obtain and can be costly when it is. This paper presents theoretical results focusing on the impact of confounding on decision making, emphasizing that optimizing decisions often involves determining whether a causal effect exceeds a threshold rather than minimizing bias in the estimate. Consequently, models built with readily available but confounded data can sometimes yield decisions as good as or better than those based on costly, unconfounded data. This can occur when larger effects are more likely to be overestimated or when the benefits of larger, cheaper data sets outweigh the drawbacks of confounding. We validate the theoretical findings using benchmark data from the 2016 Atlantic Causal Inference Conference causal modeling competition, encompassing 77 scenarios and 7,700 data sets. We then introduce theoretical conditions, weaker than ignorability, that characterize when confounding preserves effect rankings. These conditions allow for empirical heuristic tests to assess whether observational data aligns with this structure. Finally, we apply our findings in a large-scale case study using advertising data, demonstrating how these insights can guide decision making in practice.

Funding: This research, including Yanfang Hou's contributions, was supported by the Research Grants Council [Grant 26500822]. The authors thank Ira Rennert and the New York University/Stern Fubon Center for support.

Data Ethics & Reproducibility Note: The code capsule is available on Code Ocean at https://doi.org/10.24433/CO.6587526.v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2023.0012).

Keywords: causal inference • causal decision making • observational data • confounding bias • experimental data • treatment assignment

1. Introduction

Statistical models for causal estimation are increasingly being used to inform intervention decisions. Applications range from advertising (Radcliffe and Surry 2011) and customer retention (Ascarza 2018) to the allocation of subsidies for key resources (Bhattacharya and Dupas 2012) and personalized medicine (Kent et al. 2020). The primary role of these statistical models is to estimate how the effects of an intervention (e.g., an incentive, a medical treatment) vary among individuals. This capability enables decision makers to allocate interventions to those individuals for whom the intervention is expected to be most effective. Unlike traditional causal inference, the focus here is on capturing the variation in causal effects among individuals—who will be more or less affected—rather than on estimating the effects themselves.

The statistical (predictive) models output scores indicating the relative effectiveness of an intervention for each individual (e.g., the impact of an ad on purchases). Commonly, interventions are then allocated to individuals with the highest scores to maximize causal impact. Importantly, it does not matter if these scores correspond to actual, accurate effect estimates; the key is that individuals with the highest scores experience the largest effects.

A major implication of this distinction is that we should rethink how we handle confounding bias in predictive models that estimate heterogeneity in causal effects. In causal inference, it is standard practice to model effects so that any observed correlation between the treatment and the outcome can be attributed to a causal effect. Confounding is typically seen as something to be eradicated because it obscures the true causal relationship, meaning that the observed correlation may not accurately reflect the actual effect.

However, inferring causation is not necessarily out of the question, for example, if confounding bias is greater for individuals with larger effects. Even if we cannot estimate precise effects for each individual, we may still infer that the effect is larger for certain individuals compared with others.

Fernández-Loría and Provost (2022b) highlight the observation, originally presented in an earlier working version of this paper, that causal effect models built from confounded data can potentially lead to decisions that are as effective as those based on unconfounded data. This paper provides the theoretical analysis and insights that underpin the results reported by Fernández-Loría and Provost (2022b) and further explores the implications for causal modeling and data collection.

These are this paper's contributions:

1. We support the observation in Fernández-Loría and Provost (2022b) by theoretically demonstrating why confounding bias may not harm decision making. We achieve this by using the analytical framework developed by Fernández-Loría and Provost (2022a), who show a different result: models that entirely forgo causal modeling can still capture effect variation and thereby lead to effective decision making. We repurpose the framework to explore the bias–variance implications of training data choice in causal modeling, focusing on the setting in which bias arises from confounding and variance arises from data size. We compare models learned from confounded-but-larger data with those from unconfounded-but-smaller data, showing conditions under which decisions from confounded data models are as good as or better than those from unconfounded data models. These conditions include scenarios in which confounding pushes estimations in the correct direction (e.g., overestimating positive effects) and in which large observational data sets mitigate confounding's detrimental effects by reducing variance-related errors.

2. We empirically demonstrate that running experiments to collect unconfounded data may not always be the best alternative even in the presence of substantial confounding in observational data. This conclusion is based on an experimental comparison of confounded and unconfounded models using a large suite of benchmark data sets from a causal modeling competition at a top causal inference conference. There are two main reasons for this result. First, experimental data may not necessarily lead to better decisions because of the bias–variance trade-off. Second, even if it does, the improvement may not justify the cost of collecting the experimental data.

3. We introduce theoretical conditions that characterize when confounding bias does not distort the ranking of treatment effects. These conditions describe cases in which confounding shifts effect estimates but preserves their relative ordering, meaning that decision rules based on ranking remain valid. This insight provides a conceptual foundation for understanding when confounded models can still be effective for treatment prioritization even when unbiased effect estimation is not possible. Importantly, these conditions are weaker than the common assumption of ignorability (Rosenbaum and Rubin 1983); whereas ignorability enables unbiased estimation of treatment effects, our conditions ensure that the relative ranking of (individuals by their) effects is preserved despite bias.

4. We provide practical tools for assessing when confounded models remain useful for decision making. When we have some experimental data, uplift curves and learning curves help determine whether to rely on a confounded model, use an unconfounded model, or invest in additional unconfounded data. For situations when there is no experimental data, we introduce empirical checks to assess whether the data-generating process preserves effect rankings despite confounding, helping practitioners determine whether a confounded model can still provide meaningful insights for decision making.

Together, these contributions give a different perspective from the common view that causal inference is not possible in the presence of confounding. We hope that a clear understanding of this phenomenon—that confounding may not hamper the ability to estimate how effects vary from one individual to another—will help improve research on machine learning for causal estimation.

The structure of the paper is as follows. In Section 2, we review the challenges of estimating causal effects for making intervention decisions and discuss related work. In Section 3, we present the theoretical analysis of the causal bias–variance trade-off that may result in confounded models making better decisions than unconfounded models. Section 4 extends this analysis experimentally to a large number of benchmark data sets and also incorporates the costs of acquiring unconfounded data by conducting controlled experiments. Section 5 introduces theoretical conditions that characterize data-generating processes for which confounding bias preserves effect rankings, identifying settings in which confounded models remain useful for decision making.

Following this, we present practical tools for assessing when confounded models are appropriate in two scenarios: one in which experimental data cannot be collected and one in which it can. In the absence of

experimental data, we propose empirical checks to assess whether the data-generating process preserves effect rankings despite confounding (Section 5.2). When experimental data are available, we introduce tools such as uplift curves and learning curves to compare models and evaluate how performance improves as more unconfounded data becomes available. All these approaches are demonstrated in the context of an advertising case study in Section 6. Finally, Section 7 discusses the practical implications of our findings and outlines directions for future research.

2. Causal Estimation

We use potential outcomes to frame causality (Rubin 1974). This framework defines interventions in terms of intervention levels, which may range from two (e.g., to represent absence or presence of a treatment) to many more (e.g., to represent different prices). The framework also assumes the existence of one potential outcome associated with each intervention level for each individual, which represents the value that the outcome would take if the individual were to be exposed to the associated intervention level. Causal effects are defined in terms of the difference between two potential outcomes, which typically are framed as treatment and control.

Because we cannot expose the same individual to several intervention levels, we cannot observe more than one potential outcome, making individual causal effects unobservable. Instead, we can estimate aggregated effects by comparing groups of similar individuals who were exposed to different intervention levels. However, this requires us to make the nontestable assumption that the individuals who received the intervention (the treated) and the individuals who did not receive the intervention (the untreated) are comparable in all aspects related to the outcome with the exception of the intervention itself.

This assumption—also known as ignorability (Rosenbaum and Rubin 1983), the back-door criterion (Pearl 2009), and exogeneity (Wooldridge 2015)—is critical for the unbiased estimation of causal effects. If, for example, the treated were to have a better outcome than the untreated even without the intervention, the estimation would suffer from an upward bias; we attribute to the intervention a positive causal effect that, at least in part, should instead be attributed to another systematic difference between treated and untreated.

A common solution to satisfy this assumption is to conduct controlled experiments to collect data. Randomization ideally ensures that, on average, treated and untreated groups are comparable in all aspects other than the intervention. However, experiments can be infeasible because of political, ethical, business, or other practical reasons. Even when possible, experiments can be expensive because of intervention and opportunity costs.

Throughout this study, we use the term "confounding bias" to refer to the statistical bias[1] that results from the data-collection process as opposed to bias that results from not selecting the appropriate functional form for the statistical model. Thus, confounding bias corresponds to the bias that is produced by the violation of ignorability.

Such bias is often regarded as the kiss of death in causal inference because it normally cannot be quantified from data alone, so no additional amount of data can help if the goal is to estimate the causal effect accurately. In fact, with more data, the estimate converges to a biased (and so arguably wrong) estimate of the effect. Nevertheless, we show that confounding bias may not hamper our ability to assess how the effect varies from one individual to another.

2.1. Heterogeneous Causal Effects

As previously mentioned, statistical inference is used to estimate aggregated rather than individual causal effects. The most common aggregated effect is the average treatment effect (ATE), which represents the average causal effect in a well-defined population. However, caution is warranted when using the ATE to make intervention decisions, particularly in the presence of heterogeneous treatment effects (HTEs).

HTEs are defined in terms of the degree to which an intervention can have different causal effects on different individuals in a population. For instance, an ATE may be zero even though all individual causal effects are nonzero. This occurs if the causal effects of different individuals in the same population cancel each other out; some individuals actually benefit from the intervention, whereas others are hurt by it. In this case, using the ATE to make intervention decisions is suboptimal. Furthermore, the ATE does not discriminate at all between the individuals in the population, so it does not support making fine-grained intervention decisions (e.g., whom to treat).

One can account for HTEs through the estimation of conditional average treatment effects (CATEs). A CATE corresponds to the average causal effect conditioned on a set of available features. Thus, to the extent that individuals in the population differ on their features—and those features are related to causal effects—we may estimate different causal effects for each individual.

Of course, the assumptions discussed above must hold for these estimates to be unbiased, and individual causal effects may still differ between individuals sharing the same features. However, the estimation of HTEs

by using CATEs is a substantial improvement over using the ATE for the purposes of deciding on individual interventions.

The use of machine learning methods for estimating CATEs is widely studied. Prominent approaches include Bayesian additive regression trees (Hill 2011), random forests (Wager and Athey 2018), and regularized support vector machines (Imai and Ratkovic 2013). Our results are independent of the specific machine learning method used for effect estimation.

2.2. Causal Decision Making

We focus on the causal task of identifying individuals within a given population for whom the causal effect of an intervention is the largest. Models designed for this task can support a wide range of decision-making problems with treatment assignment being among the most prominent.

In treatment assignment problems, each possible action corresponds to a different treatment (e.g., target versus don't target). This paper examines the treatment assignment problem defined by Manski (2004), in which each individual should ideally be assigned to the treatment that yields the most beneficial potential outcome (e.g., the highest profit). We specifically consider binary treatments, in which the choices are to act or not to act.

These treatment assignment problems can be conceptualized as a threshold model, in which assignments depend on whether a predictive model's output (e.g., the treatment effect estimate) exceeds a certain threshold. For instance, we might decide to treat individuals with positive effect estimates or, under budget constraints, decide to treat the top 10% with the largest estimates. We discuss the implications of our results for other types of decision making in Section 7.2.

Although treatment assignment and causal effect estimation are distinct problems (Hirano and Porter 2009), the estimation of CATEs is central to many recent advances in treatment assignment. For example, in the econometrics literature, Bhattacharya and Dupas (2012) show how to estimate treatment assignment policies based on causal effects when there are budget constraints and illustrate their method in the context of efficient provision of antimalaria bed net subsidies. Uplift modeling (Kane et al. 2014), which consists of assigning treatments based on CATE estimates, is also used in targeting applications, such as online advertising and customer retention (Radcliffe and Surry 2011). Other researchers propose machine learning approaches that can be used to learn optimal treatment assignment regimes from data without first modeling causal effects (Zhang et al. 2012, Zhao et al. 2012, Kitagawa and Tetenov 2018).

However, machine learning methods for treatment assignment typically assume that there is no (unobserved) confounding in the data. This is worrying because modern business information systems often are designed to confound the data they produce. For example, advertisers use machine learning models to target likely buyers with ads, and websites often recommend to their users the products that they are more likely to choose. If we were to use the data produced by these systems to estimate the causal effects of ads or recommendations, the estimates will almost certainly be biased because those targeted with interventions are the ones deemed likely to have a positive result (even without an intervention). Evidence from large-scale experiments suggests this is likely to be the case even after controlling for observed confounding (Gordon et al. 2019). But what if the decision making were just as good despite evident confounding?

Answering this question is crucial because correcting for confounding can be prohibitively expensive or even impossible in many situations, forcing us to make decisions using the available data. Traditionally, observational studies for causal estimation have been downplayed in favor of controlled experiments because of the latter's ability to control for unobserved confounders. Thus, experiments are often seen as the natural approach to avoid confounding. However, as mentioned earlier, experiments are not always feasible, making it important to understand when we can infer variation in causal effects to inform decision making despite confounding. Moreover, even when experiments are feasible, they can still be costly. Therefore, understanding how confounding affects intervention decisions is vital to determine the profitability of investing in experimental data.

2.3. Closely Related Work

The distinction between causal effect estimation and causal decision making has been examined in depth by Fernández-Loría and Provost (2022b). One of the key points of that paper is that confounding may not always be detrimental to decision making; in fact, it might help identify the individuals for whom the intervention is most effective. However, although the paper referenced an earlier version of this work to support the argument, it did not provide theoretical or empirical backing for the claim. In this paper, we offer both in Sections 3 and 4.

Moreover, this paper delves deeper into the comparison between observational and experimental data for treatment assignment. We consider the return on data investment (Section 4.3) and characterize data-generating processes by which effect estimates preserve the ordering of the actual effects despite confounding bias (Section 5).

Furthermore, we propose practical tools for practitioners to assess the plausibility of using a confounded model to estimate variations in effects, encompassing scenarios in which unconfounded data cannot be acquired (Section 5.2) and in which it can (Section 6).

As part of our theoretical analysis, we explore the bias–variance trade-off in estimating causal effect models for decision making when using confounded-but-larger training data compared with unconfounded-but-smaller data. This analysis repurposes the analytical framework presented by Fernández-Loría and Provost (2022a). A key distinction is that, whereas Fernández-Loría and Provost (2022a) focus on the use of noncausal models to capture variations in causal effects, we focus on the use of causal effect models based on confounded data. In their context, bias arose from completely ignoring counterfactuals, and variance resulted from estimating multiple counterfactuals simultaneously. In our context, bias arises from confounding, and variance results from the data size. Stated differently, Fernández-Loría and Provost (2022a) examine bias and variance induced by the choice of target variable (the outcome versus the causal effect), whereas this paper examines bias and variance induced by the choice of training data.

This distinction between the sources of bias and variance means that the two studies have different implications for statistical modeling in terms of how we look at data investment, the characterization of settings in which bias might be beneficial, the assessment of potentially biased models, and the design of algorithms for treatment assignment. Our contribution regarding the bias–variance trade-off lies in understanding the implications of the interplay between confounding bias and data size and how this interplay impacts decision making; it does not lie in the analytical framework, which is the same as that used by Fernández-Loría and Provost (2022a).

3. Bias and Variance in Causal Estimation

Let τ be the optimal threshold for deciding whether to intervene on an individual based on the expected effect of the intervention. This threshold could be determined by the costs and benefits of intervening versus not intervening or by budget constraints that imply only a percentage of individuals should be targeted.[2] Then, the optimal decision for an individual with feature vector $\boldsymbol{x}$, denoted by $a^*(\boldsymbol{x}) \in 0, 1$, is given by

$$a^*(\boldsymbol{x}) = \mathbf{1}[h(\boldsymbol{x}) \geq \tau], \tag{1}$$

where $h(\boldsymbol{x})$ is the CATE given $\boldsymbol{x}$ and $\mathbf{1}$ corresponds to the indicator function. Also, suppose we have a classifier $\hat{a}$ that uses $\hat{h}$, an estimate of h,

$$\hat{a}(\boldsymbol{x}) = \mathbf{1}[\hat{h}(\boldsymbol{x}) \geq \tau]. \tag{2}$$

Two factors may lead to suboptimal decisions (i.e., $\hat{a}(\boldsymbol{x}) \neq a^*(\boldsymbol{x})$). One is that $\hat{h}$ is derived from a random sample D, so errors that result from the random nature of D may push estimates to the wrong side of τ. This often happens when the statistical procedure used to estimate $\hat{h}$ is too complex for the amount of training data available (i.e., overfitting), so individuals are segmented into subpopulations that are too fine-grained, resulting in $\hat{h}$ having a large variance.

Another potential source of errors is bias in $\hat{h}$, which implies that $\mathbb{E}[\hat{h}(\boldsymbol{x})|D] \neq h(\boldsymbol{x})$. Most of the machine learning literature focuses on the specific case in which the bias in $\hat{h}$ is produced by a learning procedure that is too simple to unbiasedly learn h (i.e., underfitting), such as when a linear regression is used to model a nonlinear relationship. Therefore, the traditional bias–variance trade-off typically discussed in machine learning relates to the ideal complexity in a statistical procedure given a certain amount of training data. We extend that perspective, noting that the bias in $\hat{h}$ may also come from the population from which D was sampled—such as when there is confounding—which is of critical importance when the target of the statistical estimation is causal.

3.1. Treatment Assignment Without Accurate Effect Estimation

Causal effect estimation and treatment assignment are affected differently by bias and variance in the estimation. We illustrate this by framing the former as a regression task with mean squared error (MSE) and by framing the latter as a classification task with weighted zero-one loss.

Let Y^Δ represent the treatment effect at the individual level, defined as

$$Y^\Delta = Y^1 - Y^0 = h(\boldsymbol{x}) + \epsilon, \tag{3}$$

where Y^1 and Y^0 denote the potential outcomes under treatment and no treatment, respectively. The term ϵ is a random variable that represents the idiosyncratic variation in the treatment effects not correlated with $\boldsymbol{x}$. We assume that epsilon has a conditional variance $\mathrm{Var}[\epsilon|\boldsymbol{x}] = \sigma$ and a conditional mean $\mathbb{E}[\epsilon|\boldsymbol{x}] = 0$.

The MSE with respect to causal effects is

$$\text{MSE}\ (\boldsymbol{x}) = \mathbb{E}[(Y^{\Delta} - \hat{h}(\boldsymbol{x}))^2 | \boldsymbol{x}, D], \tag{4}$$

which is minimized when $\hat{h}(\boldsymbol{x}) = h(\boldsymbol{x})$. Thus, methods for causal effect estimation are often designed to learn h by minimizing an estimate of Equation (4) over the entire population. In what follows, we focus on the evaluation of $\hat{h}$ at the individual level, so all quantities are presumed to be conditioned at a particular point $\boldsymbol{x}$ in the feature space (e.g., $h = h(\boldsymbol{x})$), and that explicit dependence is suppressed for convenience. Equation (4) may be further decomposed as (Geman et al. 1992)

$$\text{MSE} = (\mathbb{E}[\hat{h}|D] - h)^2 + Var[\hat{h}|D] + \sigma^2, \tag{5}$$

where σ^2 corresponds to the heterogeneity in causal effects that is not captured by $\boldsymbol{x}$. The other terms correspond to the squared bias and the variance in the statistical procedure that is used to estimate h. Thus, larger bias and variance generally imply worse causal effect estimates.

On the other hand, consider the cost of making decisions using $\hat{h}$ instead of h:

$$\begin{aligned}\text{Decision Error} &= \omega_1 a^* P(\hat{a} = 0|D) + \omega_2(1 - a^*)P(\hat{a} = 1|D) \\ &= \omega_1 a^* \int_{-\infty}^{\tau} p(\hat{h})d\hat{h} + \omega_2(1 - a^*)\int_{\tau}^{\infty} p(\hat{h})d\hat{h},\end{aligned} \tag{6}$$

where $p(\hat{h})$ is the probability density function of $\hat{h}$ given the random sample D, ω_1 is the cost of incorrectly withholding the treatment, and ω_2 is the cost of incorrectly giving the treatment. This quantity is also known as regret in decision theory and is equivalent to a weighted zero-one loss.

In practice, determining $p(\hat{h})$ exactly may not be straightforward because $p(\hat{h})$ depends on the statistical procedure that is used to fit $\hat{h}$. In order to gain some intuition on how bias and variance may affect decision error, we approximate $p(\hat{h})$ using a normal distribution. This is a reasonable approximation considering the estimand is an average (i.e., a conditional average treatment effect), and sample-mean distributions converge to a normal distribution under the central limit theorem. Indeed, the sampling distributions of several popular CATE estimators follow approximately normal distributions (e.g., Hill 2011, Wager and Athey 2018), and other related studies that use the same approximation show that their qualitative conclusions are still generally valid even when normality is not met (Friedman 1997, Fernández-Loría and Provost 2022a). Under the normality assumption, it can be shown that (Friedman 1997)

$$\begin{aligned}\text{Decision Error} = \tilde{\Phi}\left[\text{sign}(h - \tau)\frac{\mathbb{E}[\hat{h}|D] - \tau}{\sqrt{Var[\hat{h}|D]}}\right] \\ \cdot (\omega_1 a^* + \omega_2(1 - a^*)),\end{aligned} \tag{7}$$

where $\tilde{\Phi}$ is the upper tail area of the standard normal distribution:

$$\tilde{\Phi}(z) = \frac{1}{\sqrt{2\pi}}\int_z^{\infty} \exp\left(-\frac{u^2}{2}\right)du. \tag{8}$$

Equation (7) shows that bias and variance affect decision error very differently than MSE. As mentioned, bias is defined as the difference between $\mathbb{E}[\hat{h}|D]$ and h, and in traditional regression problems, a larger bias means worse predictions (see Equation (5)). However, as Equation (7) shows, bias only hurts decision making when it goes in the opposite direction of the correct decision: $(h - \tau) \cdot (\mathbb{E}[\hat{h}|D] - h) < 0$. In fact, bias helps when it goes in the same direction as the correct decision because it lessens errors caused by variance. In contrast, larger variance hurts when the estimation is on the correct side of the decision boundary (i.e., $(h - \tau) \cdot (\mathbb{E}[\hat{h}|D] - \tau) > 0$) but helps otherwise.

This formulation implies that assigning treatments to optimize their causal impact does not (necessarily) require accurate estimation of causal effects: the decision error can be small even when the MSE is large because of bias in the estimates. Moreover, Equation (7) shows that more data corrects the detrimental impact of bias when the estimation is on the correct side of the threshold. More specifically, as the size of D increases, $Var[\hat{h}|D]$ converges to zero, which implies that the decision error also converges to zero as long as $(h - \tau) \cdot (\mathbb{E}[\hat{h}|D] - \tau) > 0$. Therefore, it is possible to converge to optimal decisions even if effect estimations remain highly biased.

In our context, this means that we may be able to proceed with accurate causal decision making even when it is impossible (or prohibitively expensive) to gather unconfounded data. It also suggests that when such data can

be acquired, the size of the data sets available to build the statistical models is a critical factor. Let's examine that in more detail.

3.2. Trade-off: Confounded Data vs. Small Data

As mentioned earlier, one approach to estimate causal effect models is to use data generated by the day-to-day operations of the business, which we refer to as observational data. Although such data may be plentiful and inexpensive to acquire, observational data sets are likely to be confounded.

In settings in which experimentation is feasible, another alternative is to collect unconfounded experimental data by conducting controlled experiments. However, experiments are often costly because they must be carefully designed and implemented (Kohavi et al. 2009) and, more importantly, because they require random interventions, which implies making suboptimal decisions for some individuals in order to collect data. As a result, the amount of experimental data that may be acquired is often substantially smaller than the observational data already available. Note that having experimental data available from a prior A/B test or other controlled experiment may not be sufficient. Most A/B tests are run to estimate ATEs; estimating CATEs well enough to make good individual-level decisions requires substantially more data (Radcliffe and Surry 2011).

This creates a trade-off between supervised learning errors (i.e., errors caused by overfitting and underfitting) and errors that result from confounding bias. More specifically, when choosing between a model trained with larger-but-confounded data and a model trained with unconfounded-but-smaller data, we should keep in mind that, whereas the former may suffer from worse performance because of confounding, the latter may suffer from worse performance because of more supervised learning errors. As a result, a model trained with (confounded) observational data may lead to treatment assignments as good as or better than those produced by a model trained with (smaller) experimental data.

We formalize this idea as follows. Suppose we have the option to choose between two different causal effect models for selecting individuals to receive an intervention: (i) a model trained on observational data (the observational model (OM)) and (ii) a model trained on experimental data (the experimental model (EM)). Moreover, suppose that the observational data set is substantially larger than the experimental data set but suffers from confounding bias because of some unobserved factor(s) affecting both the treatment assignment and the outcome. If this is the case, how can we characterize the settings in which OM has a smaller decision error (Equation (7)) than EM?

We use the analytical framework presented by Fernández-Loría and Provost (2022a) to answer this question. This prior study also framed treatment assignments as a classification task to compare competing models but focused on a different setting: comparing a causal effect model and a simple outcome prediction model that completely ignores the question of causal influence (e.g., a traditional predictive model that only estimates how likely someone is to make a purchase but is used to decide whom to target with advertisement). In that context, bias and variance came from the exclusion versus inclusion of explicit causal (i.e., counterfactual) estimation.

The following analysis adapts the framework of Fernández-Loría and Provost (2022a) to settings in which both competing models estimate causal effects. In this context, the bias arises from confounding, whereas the variance is due to the size of the data used to estimate $\hat{h}$.

3.3. Implications of Confounding for Decision Making

Let $\hat{h}_{OM}$ be OM's estimate and $\hat{h}_{EM}$ be EM's estimate. We consider next scenarios in which confounding is the sole source of bias, meaning there is no bias in the modeling procedures used to learn $\hat{h}_{OM}$ and $\hat{h}_{EM}$. An extended analysis in the appendices addresses cases in which model bias is also present. In the absence of bias in the modeling procedures, only $\hat{h}_{OM}$ is biased because of confounding, whereas $\hat{h}_{EM}$ remains unbiased.

Theorem 1. *Let h be the average treatment effect given a specific set of feature values, and let $\hat{h}_{OM}$ and $\hat{h}_{EM}$ be the effect estimates provided by two causal effect models. Assume $\hat{h}_{EM}$ is unbiased. If $\hat{h}_{OM}$ and $\hat{h}_{EM}$ are normally distributed (conditional on the set of feature values), then $\hat{h}_{OM}$ leads to lower decision error than $\hat{h}_{EM}$ for individuals with those feature values if*

$$\frac{b}{m} < 1 - \sqrt{\gamma}, \tag{9}$$

where $b = h - \mathbb{E}[\hat{h}_{OM}]$ is the (negation of the) bias in $\hat{h}_{OM}$, $\gamma = Var[\hat{h}_{OM}]/Var[\hat{h}_{EM}]$ is the ratio of the variances of $\hat{h}_{OM}$ and $\hat{h}_{EM}$, and $m = h - \tau$ is the causal margin—the distance between the average treatment effect and the optimal decision boundary (or threshold) when h is used to make decisions.

Proofs for this and all other analytical derivations can be found in the appendices.

For mathematical clarity, Theorem 1 casts bias in relative terms, (b/m)—the causal boundary bias (to which we refer simply as boundary bias). This quantity can be interpreted as the relative distance that the estimates are pushed toward (or past) the decision boundary as a result of bias (so a larger quantity is worse). Whether OM is better than EM, thus, depends on the boundary bias produced by the confounding and the relative decrease in variance when using OM instead of EM.

Because we are specifically interested in comparing cases in which OM is trained on substantially more data than EM, we assume $\gamma \leq 1$.[3] Under this assumption, Equation (9) provides a clear criterion for determining when each approach performs better:

1. If the boundary bias is larger than one, the EM makes better decisions.
2. Otherwise, if the boundary bias is larger than zero, the EM makes better decisions only if the decrease in variance when using the OM is small compared with the boundary bias.
3. If the boundary bias is nonpositive, the OM makes better decisions.

Figure 1 provides an example for each of the three scenarios. Recall that EM is unbiased, so its distribution is centered at h in each case. OM is biased because of confounding, so its distribution is not centered at h. The vertical line at τ is the decision boundary. The probability of making the wrong decision corresponds to the area under the curves to the left side of τ.

As expected, EM is preferable if the boundary bias is large (Figure 1(a)). Considering that boundary bias is the ratio between the bias and the margin, a boundary bias larger than one implies that the bias produced by confounding is strong enough for the average estimated causal effect to be on the wrong side of the decision boundary. In this scenario, no matter how large the decrease in variance is from having more confounded data, EM always performs better.

However, if the boundary bias is positive but smaller than one (Figure 1(b)), EM performs better only if OM's decrease in variance is small compared with the boundary bias. Figure 2 shows the minimum decrease in variance that would support choosing OM. Any decrease above the curve would mean that OM is preferable. The key insight here is that, in this scenario, the larger size of the observational data can correct the detrimental effect of confounding bias because the estimate converges to the correct decision with increasing data despite being biased.

Finally, and most interestingly, if boundary bias is negative (Figure 1(c)), OM always performs better. Boundary bias is negative when the confounding bias pushes the estimations further in the direction of the correct decision, for example, when large causal effects are overestimated. This could occur if there is a correlation between the confounding bias and the causal effect, for example, if confounding is stronger for likely buyers and the effect of ads is also stronger for them. Section 5 discusses data-generating processes in which this could happen systematically.

4. Experimental Analysis

The previous section shows theoretically that there are regimes in which confounded models can lead to better decisions than unconfounded models (and vice versa). But do these regimes actually occur? In this section, we show empirically that there indeed are many settings in which confounded models can work at least as well as unconfounded ones.

To compare confounded models and unconfounded models across multiple settings, we use the suite of data sets introduced at the 2016 Atlantic Causal Inference Conference (ACIC) for comparing causal estimation

Figure 1. (Color online) Sampling Distributions of $\hat{h}$ When EM Is Unbiased

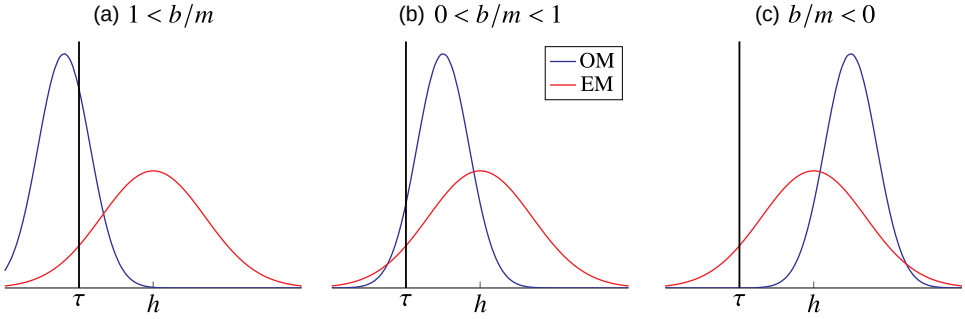
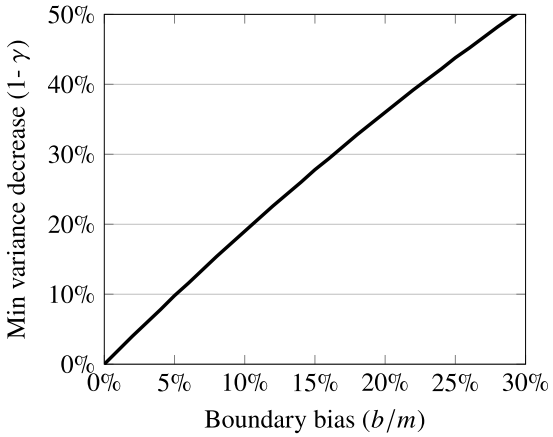


Figure 2. Minimum Variance Decrease to Prefer OM over EM



methods in the presence of heterogeneous treatment effects. The data sets[4] were generated under 77 different settings varying in their level of nonlinearity (of the assignment mechanism and the response surface), the level of treatment effect heterogeneity, the ratio of treated to control observations, the overlap between treatment group and control group, the number of confounders, and the magnitude of the treatment effect. For each of the 77 settings, treatment and response data were simulated 100 times using real-world data corresponding to 4,802 individuals and 58 features. Thus, there are 7,700 data sets in total with 4,802 observations each.

4.1. Experimental Setup

We use the ACIC data sets to compare the decision-making performance of models built with experimental and observational data. The treatment assignment in the original ACIC data sets is nonrandom, corresponding to an observational study. However, the data sets were generated such that all confounders are measured, meaning there is no unobserved confounding.

We introduce unobserved confounding by adjusting the baseline outcomes for treated individuals so that they have systematically different outcomes from untreated individuals in the absence of treatment. This introduction of unobserved confounding does not alter other characteristics of interest in the data sets (e.g., magnitude of effects, level of heterogeneity).

Let T represent the treatment assignment ($T = 1$ if treated, $T = 0$ if untreated), Y^j the potential outcome under treatment assignment j, β the average effect in the population (i.e., $\beta = \mathbb{E}[Y^1 - Y^0]$), and α the (relative) confounding bias to be introduced. We add $\alpha\beta$ to the potential outcomes of all treated individuals. This results in the same treatment effects for everyone but introduces systematic differences between the potential outcomes of treated and untreated individuals. Moreover, α can be interpreted as the percentage by which the estimated average effect ($\hat{\beta}$) is biased as a result of unobserved confounding. Therefore, in the observational data, we have

$$\mathbb{E}_X[\mathbb{E}_Y[Y|\boldsymbol{X}, T = 1] - \mathbb{E}_Y[Y|\boldsymbol{X}, T = 0]] = \beta(1 + \alpha). \tag{10}$$

To generate experimental data, we replace the treatment assignment in the original data with a Bernoulli random variable that preserves the ratio of treated to control. The ACIC data sets include both potential outcomes for each individual, so we can simulate the results of any experiment.

Decision-making performance is assessed by splitting each data set into a training set (D_{tr}) of 4,000 observations and a test set (D_{te}) of 802 observations. We train a causal effect model ($\hat{h}$) using the training set and use this model to decide which individuals in the test set would benefit from the intervention. Given the set of features $\boldsymbol{x}_i$ for individual i, the intervention decision is

$$\hat{a}(\boldsymbol{x}_i) = \mathbf{1}[\hat{h}(\boldsymbol{x}_i|D_{tr}) > 0], \tag{11}$$

where $\hat{a}(\boldsymbol{x}_i) = 1$ if the decision is to treat individual i, and $\hat{a}(\boldsymbol{x}_i) = 0$ otherwise.[5]

The average effect of deploying a treatment assignment policy $\hat{a}$ is given by

$$\beta_{\hat{a}} = \mathbb{E}[\hat{a}(\boldsymbol{X})Y^{\Delta}]. \tag{12}$$

We evaluate model performance using an unbiased empirical estimate of Equation (12) in the test set, denoted as P. Specifically, let μ_i^j be the conditional expectation of the potential outcome Y^j for individual i (i.e., $\mu_i^j = \mathbb{E}[Y^j|\boldsymbol{x}_i]$). Then, P is given by

$$P = \frac{1}{|D_{te}|}\sum_{i \in D_{te}} \hat{a}(\boldsymbol{x}_i)(\mu_i^1 - \mu_i^0). \tag{13}$$

The values μ_i^1 and μ_i^0 are available in the ACIC data sets, and $|D_{te}|$ is the size of the test set.

This performance measure implies that our analysis focuses on comparing OM and EM at the population level rather than the subset level. However, it is important to note that a policy performing best on average can still underperform for certain subsets. For example, although confounding might be beneficial overall, it could be detrimental to specific groups. Conversely, an unconfounded model could perform better on average but worse for certain subsets, especially those with limited data, leading to higher variance estimates.[6]

In all the analyses that follow, we estimate $\hat{h}$ using causal forests (Wager and Athey 2018) without the honesty criterion as we found that it hurt decision-making performance in most cases. Essentially, the honesty criterion separates the training set into two samples: one to learn the tree structure and another to estimate the CATEs at each leaf. Although this ensures that causal effects are unbiasedly estimated at the leaf level, it does not imply that the predictions are better than when the entire training set is used to both learn the tree structure and estimate the leaves as is commonly done when fitting tree-based models to predict outcomes.

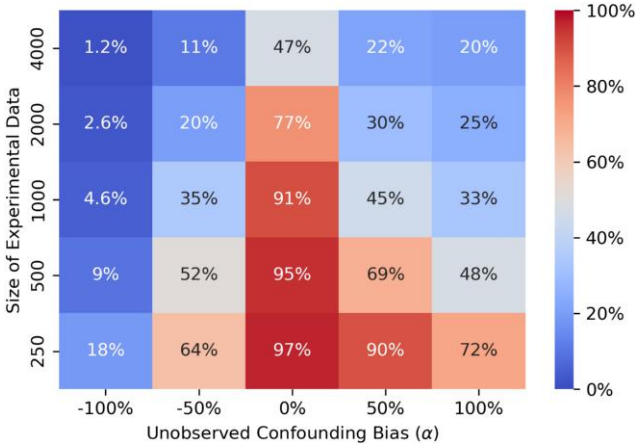
The honest approach comes at the expense of (i) more noisy estimates at the leaf level and (ii) potentially worse splits when learning tree structures as a result of using less training data to estimate the CATEs and to learn the forest, respectively. In most of the data sets in this study, fitting trees without the honest approach generally leads to better decisions even though the estimates at the leaf level are no longer unbiased. The appendix includes results of the analysis when the honesty criterion is incorporated. The analysis shows that incorporating the honesty criterion generally worsens the performance of both approaches, and—with respect to this paper's main question—it typically increases the advantage of OM relative to EM.

4.2. Comparison of Models Trained with Observational and Experimental Data

We compare the decision-making performance of models built with experimental and observational data, using different data sizes for the experimental setting and varying degrees of unobserved confounding for the observational setting. This comparison aims to analyze the causal bias–variance trade-off between observational and experimental data (as presented in Theorem 1) when many instance decisions are considered. For this analysis, we use all 7,700 ACIC data sets.

Figure 3 presents a heat map showing the percentage of data sets in which the model trained with observational data (OM) makes decisions at least as good as those made with the model trained with experimental data (EM). Specifically, it reports the percentage of data sets in which $P_{OM} \geq P_{EM}$, corresponding to the performance

Figure 3. (Color online) Percentage of Data Sets in Which OM Performs at Least as Well as EM



measure in Equation (13) when OM and EM inform the treatment assignments. The observational models were trained using 4,000 observations with unobserved confounding levels ranging from −100% underestimation to +100% overestimation.[7] The experimental models were trained with data sizes ranging from 250 to 4,000 observations.

The figure presents a clear illustration of the causal bias–variance trade-off. Notably, and in line with the theoretical results, the observational approach is often at least as good as the experimental approach when the unobserved confounding bias is not extreme and the experimental data are limited. On the other hand, the experimental approach is largely better when the confounding bias is extreme (and negative) or the experimental data set is relatively large. Therefore, whether the experimental data lead to better decisions depends on the amount of data (because of supervised learning errors) and the magnitude of the confounding (because of confounding bias errors).

Figure 3 also shows that the direction of the confounding bias is important. A positive value for α shows that the magnitudes of the effects estimated by OM are overestimated because of confounding bias, whereas a negative value implies that the magnitudes are underestimated. For example, the ATE estimate is expected to be twice as large as the actual ATE when $\alpha = 100\%$, whereas the estimate is expected to be zero when $\alpha = -100\%$. Notably, OM is more likely to perform as well as EM when $\alpha > 0$ because, as Theorem 1 shows, then the confounding bias pushes the estimations toward the decision that is best on average. As a result, the confounding bias is less likely to hurt decision making when $\alpha > 0$ and may even help lessen errors that result from variance.

However, confounding bias can still hurt decision making when $\alpha > 0$. Specifically, a positive value for α can lead to incorrect decisions for those individuals who do not behave like the average. As an example, suppose that most individuals in the population benefit from the treatment, but there also are individuals who are hurt by the treatment. Then, a large value of α may lead us to incorrectly treat individuals who should not be treated.[8] Indeed, Figure 3 shows that OM performs better when $\alpha = 50\%$ than when $\alpha = 100\%$.

Finally, as shown in the top row, OM can perform as well as EM even when the confounding is substantial and the unconfounded training data set is as large as the confounded data set. Admittedly, the fraction of cases in which this happens depends largely on the nature of the confounding (e.g., OM performs as well as EM only in 1.2% of cases when $\alpha = -100\%$ and the training set sizes are the same). However, the key takeaway is that confounding may not impair intervention decisions at all even when it results in highly inaccurate causal effect estimates. Additionally, the results presented in Figure 3 do not consider the cost of acquiring the experimental data, which we address in the next section.

4.3. Accounting for Investment Costs

The costs associated with experimentation can vary greatly depending on the context. For example, models might be used to target the top k individuals with the highest predictions or those whose estimated effect exceeds the cost of treatment. In some situations, the opportunity cost of not treating (e.g., a lost sale) is greater than the cost of treating, whereas in others, the opposite is true (e.g., wasted resources). Decisions can sometimes be revisited, but in other cases, they are final. The cost of experimentation differs in each of these scenarios.

We focus on a situation in which the goal is to maximize the causal impact of n treatment assignments (e.g., determining which n customers should receive a retention incentive). We assume these treatment decisions are irreversible (e.g., a customer not receiving an incentive cannot be offered one later if the customer leaves). To maximize causal impact, individuals are treated if the estimated effect is positive, using a threshold of zero as established earlier.[9]

We consider two approaches. One involves using a causal effect model trained on confounded observational data (OM). The other involves conducting a controlled experiment with m individuals and then using a causal effect model trained on the resulting unconfounded experimental data (EM) to make the remaining $n - m$ treatment assignments.

Conducting the experiment involves randomly selecting two groups: one to receive the treatment (treatment group) and one from which to withhold the treatment (control group). The main issue is that, because the groups are selected at random, investing in experimental data has an opportunity cost if OM could have been used to make intervention decisions for those individuals.[10] Thus, the cost of experimentation arises from the m individuals whose treatment assignments are less effective than if OM had been used. OM is preferable to EM when the cost of conducting the experiment exceeds the benefit of the (potential) improvement in decision making.

Let's quantify this cost. We have $m \leq n$ individuals who participate in the experiment (the experimental training set). Let β represent the ATE and θ denote the fraction of individuals in the experiment who are in the treatment group. We define β_{OM} as the average effect of deploying a treatment assignment policy informed by OM as

given by Equation (12) when assignments are based on OM. Then, the cost of the experiment is

$$\text{Cost} = m \cdot (\beta_{OM} - \beta\theta). \tag{14}$$

Nonetheless, EM may lead to better intervention decisions than OM for individuals who did not participate in the experiment (the inference set). Let β_{EM} be the average effect of a treatment assignment policy informed by EM as defined in Equation (12). The benefit of the experiment is

$$\text{Benefit} = (n - m) \cdot (\beta_{EM} - \beta_{OM}), \tag{15}$$

and EM is a better alternative than OM if

$$\text{Cost} < \text{Benefit}. \tag{16}$$

A key aspect when considering costs and benefits is the number of intervention decisions that could possibly be made: n. As we increase the size of the experimental training set, in order to improve EM, we decrease the number of individuals who benefit from EM (the size of the training set is m, and the size of the inference set is $n - m$). Thus, Equation (16) can be rearranged to define the break-even ratio (BER), which determines the minimum size that the inference set should have for the experiment to be profitable. Specifically, the experiment is profitable when

$$\text{BER} = \frac{\beta_{OM} - \beta\theta}{\beta_{EM} - \beta_{OM}} < \frac{n - m}{m}. \tag{17}$$

Therefore, the BER defines the minimum inference set size for EM to provide better results than OM. For example, a BER of three would imply that running the experiment and training a model with the resulting experimental data leads to better decisions than using the observational data only if the inference set size is at least three times larger than the size of the experiment.

We now compare OM and EM by estimating the BER in the ACIC data sets. For each data set, we estimate β_{OM} and β_{EM} using P as defined in Equation (13). Moreover, for each data set, we estimate β and θ using both D_{tr} and D_{te}.

Our analysis focuses on data sets in which investing in experimental data improves decision making (positive benefit) and the observational model outperforms random decisions (positive cost). We exclude data sets in which the benefit or the cost is nonpositive as the BER is inconsequential in these cases. Specifically, when the benefit is nonpositive, it indicates that investing in the experiment does not improve decision making, making OM preferable (first row of Table 1). Conversely, when the cost is nonpositive, it suggests that the observational model leads to decisions worse than treating at random, making EM preferable (second row of Table 1). The number of data sets left for this analysis—that is, those for which the choice depends on the specific costs and benefits—is shown in the third row of Table 1.

We show the results of our analysis in Figure 4. The box plots show the distribution of the BER under various degrees of confounding. The insight we get from the figure concurs with that from the previous section: the BER tends to be smaller when the magnitude of the bias is large, particularly when the bias pushes the causal effect estimates in the opposite direction of the decision that is optimal on average. However, even in the extreme case in which the confounding cancels out the average effect ($\alpha = -100\%$), more than 25% of the data sets have a BER greater than four. In these cases, conducting an experiment with 4,000 individuals to collect data and train EM leads to worse overall results than using OM if there are fewer than 16,000 individuals left to apply the model.

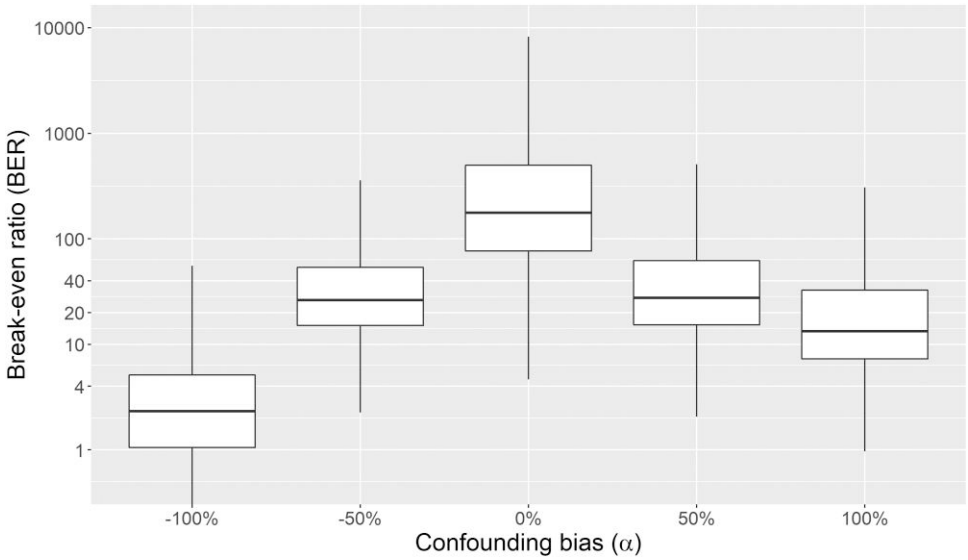
This gap is even more noticeable when the confounding bias is not as extreme. For instance, when α is -50% or 50%, the median of the BER is around 30. Therefore, in these scenarios, OM leads to better results than EM in approximately 50% of the data sets if the inference set size is smaller than 120,000. Even in the case in which there is an overestimation of 100%, the figure shows that the median of the BER is larger than 10; thus, OM is also likely to be preferable in extreme cases if the inference set is not large enough. (And recall that this analysis is for the subset of the data sets in which OM is not simply always preferable as reported in Table 1.)

Table 1. Categorization of Data Sets for BER Analysis

	Confounding bias (α)				
	−100%	−50%	0%	+50%	+100%
Data sets with nonpositive benefit (choose OM)	96	839	3,641	1,671	1,564
Data sets with nonpositive cost (choose EM)	21	8	4	2	1
Data sets in the break-even ratio (BER) analysis	7,583	6,853	4,055	6,028	6,136

Notes. Only the data sets in the third row are analyzed in Figure 4. The BER is inconsequential for the data sets in the first and second rows.

Figure 4. Distribution of Break-Even Ratio Under Various Degrees of Confounding (α)



5. Inferring Causal Variation in the Presence of Confounding

Confounding presents a major challenge for causal inference. Conventional wisdom suggests that confounded estimates should be disregarded entirely because they do not represent causal effects. However, as our empirical and theoretical analyses show, this view is overly restrictive. Even if confounding biases causal estimates, these estimates can still be useful for decision making if they help distinguish between individuals with low and high treatment effects. In other words, despite the bias, they may still capture meaningful variation in causal effects.

The key question, then, is not whether a confounded model provides unbiased estimates but under what conditions can it still reveal variation in causal effects. To address this, we characterize data-generating processes in which confounded estimates preserve the relative ordering of true causal effects. Instead of relying on the (strong) ignorability assumption, we introduce a weaker set of assumptions under which variation in causal effects can still be inferred. These assumptions do not enable unbiased estimation but nevertheless allow us to determine which individuals are likely to experience larger or smaller effects.

Of course, as with ignorability, these assumptions are ultimately unknowable in practice. When experimental data are available, there may be little need to test whether they hold; the key question is simply whether decisions based on the observational model lead to more effective outcomes regardless of whether the assumptions are valid. In Section 6, we show how to empirically evaluate an observational model in such cases.

However, when no experimental data are available, the causal inference literature offers little guidance on how to proceed when causal estimates are known to be confounded. To address this, we propose heuristic tests grounded in these assumptions to assess whether observational data suggests that effect variation can still be inferred despite bias. These tests provide a practical tool for assessing when a confounded model remains useful even in the challenging scenario in which no experimental data are available.

5.1. When Does Confounding Preserve Effect Variation?

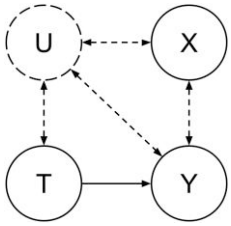
Confounding can bias effect estimates, but it does not necessarily prevent us from detecting variation in causal effects. The key question is how much it distorts the relative ordering of individual effects. To analyze this, let's model treatment assignment using an unobserved latent variable U that represents (everything related to) the propensity for treatment. Formally, in this model treatment is assigned when

$$T = \mathbf{1}[U > 0]. \tag{18}$$

This thresholding process is commonly used in discrete choice models (Train 2009), in which U may represent an individual's utility for treatment and, thus, affect (self-)selection into treatment. More generally, U captures any underlying factors that make treatment more or less likely, such as individual behaviors, external policies, or other assignment mechanisms.

Unobserved confounding occurs when U correlates with Y in ways not fully accounted for by the feature vector $\boldsymbol{X}$ as illustrated in Figure 5. The solid, one-headed arrow indicates a causal relationship, whereas the two-

Figure 5. Unobserved Confounding



Note. T connects to Y through U.

headed, dashed arrows represent correlation. Because U is associated with both treatment assignment and outcomes, the observed relationship between T and Y is confounded, meaning that their correlation does not solely reflect causation but also the dependence introduced by U. As a result, unbiased estimation of causal effects from the data are impossible.

To reason about the confounding, we decompose the outcome of interest into two components:

$$Y = TY^{\Delta} + Y^{0}. \tag{19}$$

Here, Y^{Δ} represents the treatment effect (that would be realized if treated), and Y^{0} represents the baseline outcome in the absence of treatment. With this decomposition, we can reinterpret Figure 5 as Figure 6, providing a clearer view of how $\boldsymbol{X}$ and U relate to Y. If $\boldsymbol{X}$ and U are correlated with Y^{Δ}, they act as moderators and are, therefore, important for estimating variations in causal effects. Conversely, if they are correlated only with Y^{0}, they are not informative of variations in causal effects but, nevertheless, introduce confounding if they are also related to T.

The standard ignorability assumption requires that treatment assignment be conditionally independent of potential outcomes given the observed features $\boldsymbol{X}$. This corresponds to no arrows from U to either Y^{Δ} or Y^{0} as shown in Figure 7. An arrow from U to Y^{Δ} would imply differential treatment effect bias, meaning the effect differs for treated and untreated individuals. An arrow from U to Y^{0} would imply baseline bias, indicating that treated individuals would have different outcomes from untreated ones even if they had not undergone treatment.

In what follows, we relax the ignorability assumption and instead introduce a weaker set of assumptions about confounding in which U is connected to both Y^{Δ} and Y^{0}. These assumptions define one class of data-generating processes in which, although unbiased estimation is not possible, the ranking of individuals by their causal effects can be inferred.

5.1.1. Assumption 1: Mean-Shifted Sigmoid Treatment Propensity. This assumption implies that U depends on $\boldsymbol{X}$ only through its mean, and conditional on $\boldsymbol{X}$, U has a log-concave distribution with a sigmoid-shaped cumulative distribution function (CDF). Specifically, we define U as

$$U = \mu(\boldsymbol{X}) + \xi, \tag{20}$$

Figure 6. Y Is Decomposed into the Effect Y^{Δ} and the Baseline Outcome Y^{0}

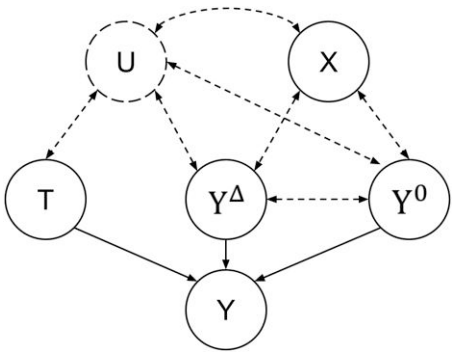
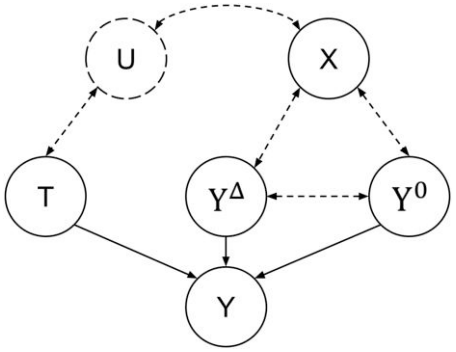


Figure 7. Ignorability



Note. There is no arrow from U to either Y^{Δ} or Y^{0}.

where the random variable ξ has a probability density function (PDF) f that satisfies the following:

$$\xi \perp \boldsymbol{X},\ \mathbb{E}[\xi] = 0, \tag{21}$$
$$f \text{ is log concave}, \tag{22}$$
$$f(u) = f(-u) \text{ for } u \in (-\infty, \infty), \tag{23}$$
$$f'(u) > 0 \text{ for } u \in (-\infty, 0), \tag{24}$$
$$f'(u) < 0 \text{ for } u \in (0, \infty). \tag{25}$$

These conditions imply that the error term ξ has a symmetric, unimodal, and log-concave PDF, resulting in the treatment propensity $U|\boldsymbol{X}$ having a sigmoid-shaped CDF. Common distributions satisfying these properties include the normal and logistic distributions, frequently used in discrete choice modeling (Train 2009).

5.1.2. Assumption 2: Linear Dependence on Confounder.

$$\mathbb{E}[Y^0|U, \boldsymbol{X}] = \alpha_0^b + \alpha_u^b U + \alpha_x^b(\boldsymbol{X}), \tag{26}$$
$$\mathbb{E}[Y^{\Delta}|U, \boldsymbol{X}] = \alpha_0^{\Delta} + \alpha_u^{\Delta} U + \alpha_x^{\Delta}(\boldsymbol{X}). \tag{27}$$

This assumption implies that both the baseline outcome and treatment effect have a linear relationship with the unobserved confounder U with no interactions between U and $\boldsymbol{X}$.

5.1.3. Assumption 3: Consistent Moderation.

$$\mathbb{E}[U|\boldsymbol{X} = \boldsymbol{x}_i] - \mathbb{E}[U|\boldsymbol{X} = \boldsymbol{x}_j] > 0$$
$$\Rightarrow \alpha_u^{\Delta}(\alpha_x^{\Delta}(\boldsymbol{x}_i) - \alpha_x^{\Delta}(\boldsymbol{x}_j)) \geq 0 \quad \text{for any } \boldsymbol{x}_i, \boldsymbol{x}_j. \tag{28}$$

This assumption implies that the way features ($\boldsymbol{X}$) correlate with treatment effects is consistent across different correlation pathways. As Figure 6 shows, features can correlate with treatment effects through multiple channels: (i) directly, (ii) indirectly through the unobserved confounder (U), and (iii) indirectly through the baseline outcome (Y^0). The key requirement is that the correlation between $\boldsymbol{X}$ and treatment effects through U does not contradict its correlation through the other pathways. That is, if individuals with higher U tend to experience larger (or smaller) treatment effects, then features predictive of higher U should moderate the effect in the same direction through other pathways. This ensures that the indirect relationship between $\boldsymbol{X}$ and Y^{Δ} through U does not reverse the overall moderation pattern between features and treatment effects.

Formally, the term α_u^{Δ} captures how U moderates treatment effects, whereas α_x^{Δ} captures how features relate to treatment effects through other pathways. This assumption ensures that, if feature values $\boldsymbol{x}_i$ predict a higher treatment propensity than $\boldsymbol{x}_j$, then the direction they moderate treatment effects aligns with the direction in which U moderates treatment effects, preventing confounding from distorting the ranking of treatment effects.

Theorem 2. *Let the confounded estimand of the CATE be*

$$\mathbb{E}[Y|T = 1, \boldsymbol{X}] - \mathbb{E}[Y|T = 0, \boldsymbol{X}]. \tag{29}$$

Under assumptions 1–3, *a larger confounded estimand of the CATE also implies a larger CATE if the following inequality is true for all individuals in the population:*

$$\alpha_u^{\Delta}\alpha_u^{b}(\mathbb{P}[T=1|\boldsymbol{X}]-0.5)>0. \tag{30}$$

Therefore, whether larger confounded estimates imply larger effects depends on the direction of the differential effect bias (α_u^{Δ}), the baseline bias (α_u^{b}), and the probability of treatment assignment.

For example, based on Equation (30), we can infer variation in causal effects if

- Higher treatment propensity predicts a larger causal effect ($\alpha_u^{\Delta}>0$).
- Higher treatment propensity predicts a larger baseline outcome ($\alpha_u^{b}>0$).
- Individuals are generally likely to receive treatment ($\mathbb{P}[T=1|\boldsymbol{X}]>0.5$).

Or if

- Higher treatment propensity predicts a larger causal effect ($\alpha_u^{\Delta}>0$).
- Higher treatment propensity predicts a smaller baseline outcome ($\alpha_u^{b}<0$).
- Individuals are generally unlikely to receive treatment ($\mathbb{P}[T=1|\boldsymbol{X}]<0.5$).

5.2. Practical Application of Assumptions

Building on the previous section, we now turn to the practical question of when a confounded model remains useful for ranking individuals by treatment effect. The assumptions introduced earlier provide a foundation for answering this question, but their implications for real-world applications need further clarification. In particular, whereas some assumptions serve primarily as analytical tools, others play a crucial role in determining whether a confounded model can still provide meaningful guidance for decision making.

Among the three assumptions, assumptions 1 and 2 (mean-shifted sigmoid treatment propensity and linear dependence on confounder) serve primarily as modeling conveniences. They simplify mathematical analysis but are not expected to hold exactly in practice. Linearity is a common assumption in statistical modeling, and sigmoid-shaped treatment propensities align with widely used models of discrete choice. Because these assumptions are primarily for analytical convenience, practitioners should not need to validate them in practice.

Assumption 3 (consistent moderation), however, is central to determining whether a confounded model retains a useful ranking signal for treatment effects. Unlike the other two assumptions, it directly concerns the data-generating process and can be partially evaluated using observational data. This assumption does not need to hold exactly for the model to be helpful; what matters is whether it holds well enough to improve decision making relative to the alternative (at the extreme, having no model at all).

To help practitioners assess whether a confounded model remains useful, we introduce a set of heuristic tests—empirical checks designed to evaluate whether the consistent moderation assumption and the conditions outlined in Theorem 2 approximately hold. These tests are particularly relevant when experiments are infeasible and only observational data are available. In contrast, Section 6 discusses how to evaluate observational models when experimental data are available. Notably, these tests rely on weaker assumptions than standard ignorability, allowing treatment assignment and outcomes to depend on unobserved confounders.

Passing these tests does not guarantee that a confounded model will accurately rank individuals. Instead, their primary purpose is diagnostic: they identify when the available observational data clearly indicates that the model should not be trusted to infer meaningful variation in effects. If the data fails these heuristic checks, then practitioners should strongly consider abandoning the model altogether or, at a minimum, drastically limiting its role in decision making.

We now introduce these heuristic tests:

1. Check whether treatment propensity is skewed. Estimate propensity scores (i.e., the probability of receiving treatment) for all individuals. Using the confounded model is potentially reliable if propensity scores are consistently high (above 50%) or low (below 50%) across the population. Otherwise, the data suggest that the model may be unreliable, and practitioners should strongly consider avoiding the use of the model, particularly when ignorability is questionable.

2. Estimate the correlation between treatment propensity, effect estimates, and baseline predictions. Compute the correlation between propensity scores and effect estimates as well as the correlation between propensity scores and predicted baseline outcomes (i.e., predicted outcomes under no treatment). Although unobserved confounding cannot be directly measured, the direction of observed confounding may be informative about the direction of unobserved confounding. If this assumption holds, the signs of these correlations can provide an estimate of the signs of α_u^{Δ} and α_u^{b}. A positive correlation with effect estimates suggests that confounding leads to positive differential treatment effect bias ($\alpha_u^{\Delta}>0$), whereas a negative correlation suggests the opposite. Similarly, a positive

correlation with baseline predictions suggests positive baseline bias ($\alpha_u^b > 0$), whereas a negative correlation suggests the opposite.

3. Evaluate whether the ranking signal condition holds. Using the estimated signs of α_u^Δ and α_u^b, check whether the inequality from Theorem 2 is satisfied (Equation (30)). If the inequality is violated, it suggests that the model may not reliably rank individuals by treatment effect. In such cases, practitioners should strongly consider either abandoning the model or strictly limiting its use, particularly when ignorability is questionable. If the inequality holds, the model likely provides a meaningful ranking signal, but its reliability should still be confirmed using the final check.

4. Assess the consistent moderation assumption. To evaluate whether assumption 3 (consistent moderation) approximately holds, practitioners should check for features that influence both treatment propensity and treatment effect estimates. If these features create contradictory moderation effects, the ranking signal may be distorted. We propose the following procedure:

   a. Identify features that are strongly correlated with both treatment propensity scores and observed effect estimates. Compute the product of these two correlation coefficients.

   b. Compare the sign of this product to the estimated sign of α_u^Δ. If the product and α_u^Δ have the same sign, the feature supports the assumption of consistent moderation. If they have opposite signs, this suggests a violation of consistent moderation.

   c. If multiple features indicate violations—or if violations occur in features with strong correlations—this suggests that the confounded model may not provide a reliable ranking signal and should not be used. If only minor violations are found, the model may still be useful, but practitioners should proceed with caution.

We emphasize again that, even if all empirical checks suggest the assumptions approximately hold, this does not guarantee that confounded estimates accurately capture variation in causal effects. These heuristic tests serve primarily as diagnostic tools, identifying clear indicators of unreliability rather than certifying model correctness. Practitioners should interpret the results with caution and consider conducting additional sensitivity analyses. An example of such an analysis, applied to decisions informed by an observational model, is provided in the appendix.

6. Assessing Confounded Models for Decision Making

This section examines how practitioners can choose between using confounded models, unconfounded models, or investing in additional experimental data. We illustrate this decision-making process using a large, publicly available data set from an online advertising randomized controlled trial. Because the data set includes treatments, business outcomes, and predictive features, it allows a direct comparison of the performance of confounded versus unconfounded models in realistic decision-making contexts. We begin by introducing the data set and describing the experimental setup, followed by practical methods for evaluating these models.

6.1. Advertising Data Set and Causal Models

The data used in our case study comes from a large-scale randomized experiment conducted by the advertising platform Criteo, which randomly targeted online advertisements to millions of consumers (Diemert et al. 2018).[11] This data set was released in order to benchmark methods for uplift modeling. The data set includes nearly 14 million user observations with 11 predictive features, a treatment indicator for ad exposure, and a binary outcome indicating whether the user visited the advertised website (visit rate). The treatment assignment probability is approximately 85%, the baseline visit rate is 4.70%, and the ATE across the full sample is 1.03%. We randomly split the data set into equally sized training and evaluation sets (50% each).

To simulate a realistic scenario of observational confounding, we introduce selection bias by systematically removing observations based on the feature "f9," which is the strongest predictor of website visits. Specifically, we remove the 10% of treated users with the lowest values of "f9" and the 10% of control users with the highest values and then remove "f9" itself from the data set. This procedure simulates a common scenario in practice: unobserved confounding that results from selective targeting of advertisements toward consumers more likely to convert. As a result, the confounded data set shrinks to approximately 6.3 million users, and the naive estimate of the ATE (the simple difference in visit rates between treated and control groups) becomes severely biased, rising to 4.08%—nearly four times the true effect.

Following the methodology in Section 4, we use causal forests to estimate treatment effects and make targeting decisions. All models are trained using all features except "f9," which is excluded to introduce confounding. Although some remaining features are correlated with "f9," allowing the model to partially adjust for confounding, this adjustment is incomplete as we show next.

Figure 8. (Color online) Estimates of the ATE of the Ad on the Visit Rate

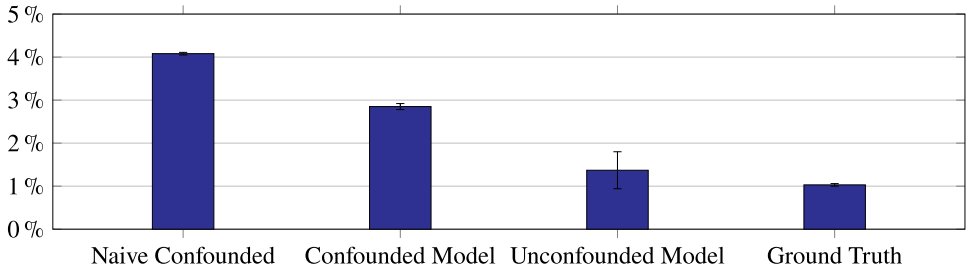


Figure 8 compares four ATE estimates: (i) a naive confounded estimate, calculated simply as the difference in visit rates between treated and untreated groups in the biased data set; (ii) an estimate from a causal forest trained on the confounded data set; (iii) an estimate from a causal forest trained on a small, experimentally collected (unconfounded) data set of 50,000 users; and (iv) the unbiased ground truth estimate obtained from the original, full data set.

The confounded model reduces bias compared with the naive estimate but still noticeably overestimates the true ATE. Conversely, the unconfounded model trained on far fewer observations (just 50,000) provides a more accurate estimate. This clearly demonstrates the benefit of even limited experimental data for estimating the overall average effect. However, as we show next, more accurate ATE estimation does not always translate into better decision making.

6.2. No Experimental Data

Before turning to scenarios with experimental data, we first evaluate the confounded model using only the observational data and applying the heuristic tests from Section 5.2. To implement these guidelines, we estimate treatment assignment probabilities (propensity scores) and baseline visit probabilities (outcomes when not treated) using random forests. We then apply the four empirical checks previously introduced:

- Skewness test: The estimated propensity scores are consistently high (from 82% to 100%), satisfying the skewness condition.
- Estimate bias signs: The correlation between propensity scores and estimated treatment effects is 0.76, suggesting positive differential effect bias. The correlation between propensity scores and the probability of a visit in the control group is 0.75, suggesting positive baseline bias.
- Theorem validation: Given the correlations and the direction of the skewness, Theorem 2 appears to hold.
- Consistent moderation: We assess whether the remaining features (excluding "f9") exhibit consistent relationships with the sign of the estimated differential effect bias. Nearly all features follow the expected pattern: they either have positive correlations with both treatment assignment propensity and effect estimates or negative correlations with both, supporting the consistent moderation assumption. The only exception is feature "f2," which shows a minor violation (negative correlation with propensity: −0.06, positive correlation with effect estimates: 0.04). However, this deviation is small and unlikely to meaningfully distort ranking signals, especially given that all other features exhibit stronger correlations with both treatment propensity and effect estimates.

These results suggest that, although the confounded model should be used with caution, it likely captures meaningful variation in treatment effects. In the next section, we evaluate with experimental data the confounded model's performance in estimating variation in causal effects.

6.3. Limited Experimental Data

We first consider settings in which limited experimental data are available, but acquiring more is not an option, which is not unusual in practice. A typical scenario would be that an experiment was run to estimate the overall effectiveness of the ad campaign (the ATE), but the resulting data may not be large enough to accurately estimate individual-level effectiveness (the CATEs). In this example, advertisers may be unwilling to randomly withhold/target ads to a large number of consumers in order to obtain enough data to learn an accurate CATE model, but they may still be willing to do so for a small number of individuals for evaluating the effectiveness of the campaign. Such data can also be used to assess the decision-making performance of a potentially confounded OM.

For our case study, suppose that the advertiser has access to experimental data from an A/B test that was run with 50,000 online consumers. We create this data set by randomly sampling 50,000 individuals from the original

(unconfounded) training data set. Such data could be the result of an A/B test that was performed to assess whether the ad has a positive effect (on average) before deploying the advertising campaign at scale. With a sample size of 50,000, the standard error for the ATE estimate is 0.24%, meaning that such a sample size would result in a well-powered A/B test with a significance level below 0.0001.

However, the fact that experimental data results in a better ATE estimate does not imply that we should use it in lieu of observational data to make individual intervention decisions. We need to compare the two approaches to see which has better decision-making performance. To this end, we compare the two approaches based on estimates of the cumulative incremental gain in visits that is achieved when some top fraction of individuals with the largest scores (causal effect estimates) are targeted. This is represented by the curves in Figure 9, which are also known as uplift curves (Devriendt et al. 2018). Uplift curves can be built using data from any A/B test, so a practitioner who has access to some experimental data can use this tool to evaluate the decision-making performance of any scoring model. In this particular example, the curves were generated from the experimental data set consisting of 50,000 individuals. The performance of the experimental model was estimated using cross-validation.

Figure 9 shows that the observational model is generally better at making decisions than the experimental model except when only a small fraction of individuals are targeted. Thus, the uplift curves show that the observational model can lead to better targeting decisions than the experimental model despite its obviously confounded and biased estimates.

In sum, an experimental data set from an A/B test used for ATE estimation also can be used to evaluate whether it would be beneficial to use a confounded model for decision making.

6.4. Investing in (More) Experimental Data

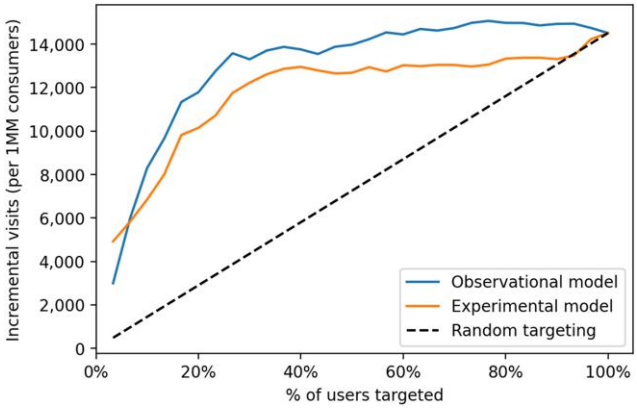
Now, let's consider settings in which one has the option to invest in additional experimental data, possibly a lot of it, at a nonnegligible cost. A primary implication of this paper is that unconfounded models are not necessarily better for decision making, so we should take seriously the question of whether the investment in unconfounded data are worthwhile. As our analysis in Section 4.3 shows, practitioners should carefully balance the cost of acquiring experimental data with the potential improvement in decision making that such data could bring. More data are not necessarily better, and the ability to run large experiments does not imply that larger experiments should be run.

As an illustration, suppose we decide to target 30% of users with advertising to maximize the return on investment (ROI) of the advertising budget given that the uplift curves in Figure 9 start to flatten beyond this point. Based on Figure 9, we know that the observational model outperforms an experimental model trained with a small amount of experimental data. However, what if we could conduct more experimentation to gather additional data? Should we?

As discussed in Section 4.3, we should invest in additional data only if the improvement in decision making is substantial enough to justify the cost.

Suppose we have n potential targets for the advertisement. One approach is to deploy the observational model and target the top 30% of individuals with the largest (confounded) effect estimates. An alternative approach is

Figure 9. (Color online) Uplift Curves for Models Trained with Larger-but-Confounded Data and Unconfounded-but-Smaller Data



to conduct an experiment with m individuals, train a model using the larger experimental data set, and then apply the new experimental model to target the top 30% of the remaining $n - m$ individuals.

Based on Equation (16) in Section 4.3, investing in more experimental data are preferable if

$$m \cdot (\beta_{OM} - \theta\beta) < (n - m) \cdot (\beta_{EM} - \beta_{OM}), \tag{31}$$

where β is the average effect, θ is the percentage of treated individuals in the experiment, and β_{OM} and β_{EM} are the average policy effects when OM and EM are used to inform treatment assignments, respectively (the mathematical definition is in Equation (12)).

Let π be the percentage improvement in decision making produced by the investment in data:

$$\pi = \frac{\beta_{EM}}{\beta_{OM}} - 1. \tag{32}$$

Based on the rearrangement of Equation (31), we should invest in more experimental data if

$$\frac{m(\beta_{OM} - \beta\theta)}{(n - m)\beta_{OM}} < \pi. \tag{33}$$

The left-hand term can be interpreted as the break-even improvement, analogous to the BER introduced in Section 4.3. To calculate it, we set n to the size of the test set (6,989,796 individuals). Given our aim to target 30% of individuals, we set $\theta = 30\%$. We estimate β_{OM} and β using the same data used to produce Figure 9 (i.e., the limited experimental data already available). Finally, we consider various values for m to account for different experiment sizes.

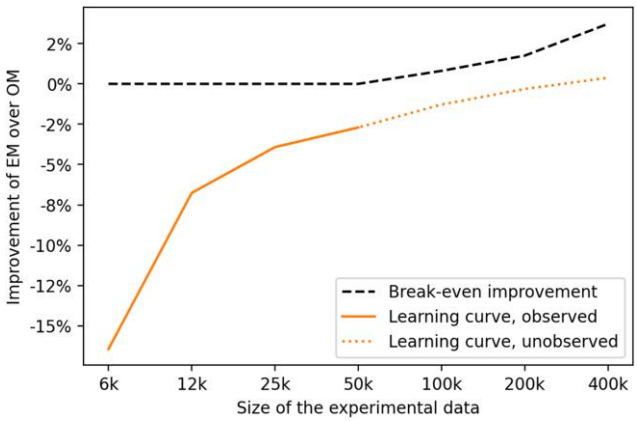
To estimate the potential ROI of the data investment, represented by π, we conduct a learning curve analysis (Provost and Fawcett 2013). Specifically, when investing in experimental data for causal decision making, we are interested in estimating whether the improvement of EM compared with OM is steep enough to surpass the break-even improvement defined in Equation (33). We illustrate this in Figure 10.[12]

The black dashed line is the break-even improvement—the left-hand side of the inequality in Equation (33). Note that the break-even improvement is zero for data sizes below 50,000 because there is already an experimental sample of that size (which means $m = 0$ in these cases). Larger data sizes would require investing in more data, so the improvement would need to be greater for the experiment to be profitable.

The orange line (the solid-and-dotted line) represents the improvement that EM could achieve with training data of various sizes, the right-hand side of the inequality in Equation (33). The solid section of the curve (the observed learning curve) is what a practitioner could observe based on the experimental data already available. The dotted section of the curve (the unobserved learning curve) is the actual rest of the learning curve, which would not be observed by a practitioner prior to running additional experimentation. When deciding whether to invest in additional experimental data, practitioners should estimate whether there is a sample size at which performance would likely surpass the competitor (Provost and Fawcett 2013), meaning in our case that the unobserved learning curve would exceed the break-even improvement.

One common method to estimate the unobserved part of a learning curve is to extrapolate based on the shape of the observed learning curve. In predictive modeling, investments in data often have diminishing marginal

Figure 10. (Color online) Learning Curve for EM vs. the Break-Even Improvement to Recover the Investment in Data



returns, making it reasonable to assume that the slope of the learning curve will decrease with more data (Frey and Fisher 1999, Perlich et al. 2003). Given the shape of the observed learning curve in Figure 10, the practitioner could reasonably infer that the learning curve would not surpass the break-even improvement with additional training data[13] and, thus, decide not to incur the expense of investing in additional experimental data and instead deploy OM. We see from the unobserved portion of the learning curve that this decision would be correct.

7. Discussion

The central insight of this study is that confounding affects causal effect estimation and causal decision making in fundamentally different ways, sometimes leading to counterintuitive situations in which using confounded data for decision models is preferable. Crucially, accurate causal effect estimates are not necessary for making effective intervention decisions. Bias and variance influence treatment assignment and effect estimation differently, and models may still converge to optimal decisions despite confounding bias. Bias can even improve decision-making performance by reducing variance-related errors when it pushes estimates further to the correct side of the decision threshold. Additionally, if the amount of available unconfounded data is significantly smaller than the confounded data set, training a model on the unconfounded data may result in worse decisions because of more supervised learning errors. Finally, even when an unconfounded model improves decision making, the benefit may not justify the cost of acquiring additional unconfounded data.

We now explore the implications of these insights for practitioners, methodological advancements, and broader decision-making contexts.

7.1. Practical Implications

These results are important because firms often have access to large amounts of observational data but limited experimental data. Running randomized experiments can be prohibitively expensive or even impossible because of the need for randomization and other logistical constraints (e.g., engineering costs, ethical concerns). As a result, practitioners and researchers frequently use observational—and, thus, potentially confounded—data to make decisions about causally affected outcomes, such as demand forecasting (Feng et al. 2014, Demirezen and Kumar 2016, Ferreira et al. 2016) and response to digital advertising (Gordon et al. 2019).

Although researchers and practitioners recognize that confounded data can be useful for decision making, it is often treated as an unfortunate necessity, used only because acquiring unconfounded data is too difficult. Our study offers a different perspective: in certain circumstances, confounded data may actually be the preferred choice. This suggests a shift in mindset: from viewing confounded data as a last resort to actively considering when and how it might improve decision performance.

In cases in which acquiring unconfounded data are impossible, practitioners face a stark choice: use a confounded model or forgo modeling altogether. Section 5 provides a framework for evaluating this decision by introducing theoretical conditions under which confounded models can still provide meaningful rankings of treatment effects. Unlike standard assumptions, such as ignorability (which enables unbiased effect estimation), our conditions focus on preserving the relative ordering of treatment effects even when confounding introduces bias. This distinction is critical because, even if a model cannot estimate treatment effects precisely, it may still be useful for decision making if it correctly ranks individuals by their relative treatment effects.

As with all fundamental assumptions in causal inference, these conditions are inherently nontestable. However, we outline empirical checks that practitioners can perform to assess whether the conditions approximately hold (Section 5.2). These checks provide a practical tool for decision makers: they do not remove confounding but offer a way to reason about whether a confounded model is likely to provide useful guidance. If these checks are satisfied, we argue that using a confounded model is preferable to using no model at all.

To further validate decision quality, practitioners with access to some unconfounded data can use it to benchmark the confounded model's performance. Section 6 illustrates how to perform this evaluation along with an analysis in Section 4.3 for determining whether acquiring more unconfounded data would be cost-effective. Other frameworks for cost-sensitive causal classification (e.g., Verbeke et al. 2023) can be adapted to weigh the trade-offs between unconfounded data collection and potential improvements in decision making.

7.2. Other Decision-Making Contexts

Our study primarily examines binary intervention decisions, in which optimal choices can be defined using a threshold so that individuals with the largest treatment effects are always the best candidates for treatment. Next, we discuss the applicability of our insights beyond this setting.

First, the key result that confounding bias does not necessarily impair decision making applies to any context in which ranking individuals by their causal effects matters more than precisely estimating those effects. In Section 5, we identify conditions under which confounded effect estimates can still preserve the correct ranking of individuals based on their true effects. This insight extends beyond threshold-based decisions and can inform any decision-making process that aims to prioritize or identify individuals with larger effects.

Our study's focus on binary decisions allows us to frame the problem as a classification task and highlight how bias and variance affect intervention decisions differently from causal effect estimation. This implies that our analytical findings do not directly apply to treatment assignments with more than two alternatives or continuous decisions, such as pricing. However, the core idea that accurate ranking is more critical than precise estimation extends to these decision settings. For example, Besbes et al. (2010) point out that, to get optimal prices, one does not need to estimate the demand function perfectly.

Specifically, our results imply that what matters when making intervention decisions is not the accurate estimation of outcomes or effects but the ranking of the alternatives available to the decision maker. Overestimating (or underestimating) the impact of an alternative typically doesn't affect decision-making performance unless it leads to ranking the best alternative below another. Consequently, confounding biases that preserve the ranking of treatment alternatives should not significantly affect nonbinary decisions either. Developing a unified framework to explain how bias and variance affect causal decision making more generally is a natural extension of this work.

Finally, our focus on threshold-based decision logic (e.g., treating all individuals with positive estimated effects or targeting the top 10%) is crucial to deriving our insights as, in these cases, treating individuals with the largest effects is generally the optimal choice for some threshold. However, in more complex decision problems—such as constrained optimization scenarios in which simple ranking-based heuristics are suboptimal (McFowland III et al. 2021)—the bias–variance trade-off may behave differently. Exploring how these effects extend to broader decision settings remains an important direction for future research.

7.3. Future Methodological Developments

We hope that this study will help inform new methods for automated decision making. One promising direction is the development of algorithms that combine multiple data sources, optimizing the trade-off between confounding bias and supervised learning errors.

Our theoretical results suggest that large-but-confounded data sets and small-but-unconfounded data sets can be combined in a way that minimizes decision-making errors. Although existing methods seek to integrate observational and experimental data for causal effect estimation (Peysakhovich and Lada 2016, Kallus et al. 2018, Athey et al. 2020, Fernández-Loría and Provost 2020, Rosenman et al. 2023), most do not explicitly focus on optimizing decision quality. Future research should explore methods that exploit confounding bias when it helps decision making, mitigating its negative effects when it does not.

In recent years, there have been substantial advances in methods for correcting confounding in observational studies. These include techniques for addressing selection bias (Yahav et al. 2016), improving the robustness of matching methods (Morucci et al. 2022), and learning decision-making policies from confounded data (Athey and Wager 2021). Typically, these approaches aim to eliminate confounding bias entirely, operating under the principle that less bias leads to better estimates.

Although this may be appropriate for causal effect estimation, we show that it may not be desirable for decision making, for which confounding can actually reduce errors by lowering variance. Our findings suggest that, rather than focusing solely on removing bias, new methods should be developed to strategically leverage confounding when it improves decision performance.

Another promising research avenue is targeted experimentation—allocating the experimental budget to regions of the feature space in which confounding most harms decision making rather than conducting large-scale randomized experiments indiscriminately. Recent work on cost-efficient experimentation (Feit and Berman 2019, Simester et al. 2020) aligns with this approach. Developing principled methods to determine when and where experimentation is most valuable could significantly improve decision-making frameworks.

Acknowledgments

The authors thank their research assistant, Yanfang Hou, for her valuable contributions to this paper. Her assistance was instrumental in proving Lemma D.2 in the appendix and in implementing some of the code for Sections 4 and 6. The authors also thank the Fubon Center and Ira Rennert for supporting research on data science at NYU/Stern.

Appendix A. Implications of Model Bias

The analysis in Section 3.3 shows that OM may lead to better decisions than EM in the absence of any model bias. In practice, however, all models learned from data are prone to model bias because optimal predictive performance is achieved by balancing underfitting (which results in errors because of model bias) and overfitting (which results in errors because of variance). In this appendix, we relax this assumption. If we incorporate bias in EM, Theorem 1 can be generalized as follows (proof in another appendix).

Theorem A.1. *Let h be the average treatment effect given a specific set of feature values, and let $\hat{h}_{OM}$ and $\hat{h}_{EM}$ be the effect estimates provided by two causal effect models. If $\hat{h}_{OM}$ and $\hat{h}_{EM}$ are normally distributed (conditional on the set of feature values), then $\hat{h}_{OM}$ leads to lower decision error than $\hat{h}_{EM}$ for individuals with those feature values if*

$$\frac{b_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma}, \tag{A.1}$$

where $b_{OM} = h - \mathbb{E}[\hat{h}_{OM}]$ is the (negation of the) bias in $\hat{h}_{OM}$, $b_{EM} = h - \mathbb{E}[\hat{h}_{EM}]$ is the (negation of the) bias in $\hat{h}_{EM}$, $\gamma = Var[\hat{h}_{OM}]/Var[\hat{h}_{EM}]$ is the ratio of the variances of $\hat{h}_{EM}$ and $\hat{h}_{OM}$, and $m = h - \tau$ is the causal margin—the distance between the average treatment effect and the optimal decision boundary (or threshold) when h is used to make decisions.

We define the (negation of the) bias in $\hat{h}_{OM}$ as

$$b_{OM} = h - \mathbb{E}[\hat{h}_{OM}] = \delta_{OM} - (\alpha + \zeta_{OM}), \tag{A.2}$$

where α is the confounding bias in $\hat{h}_{OM}$, ζ_{OM} is the model bias in $\hat{h}_{OM}$, and δ_{OM} is a constant that we subtract from $\hat{h}_{OM}$ to account for settings in which OM uses a decision threshold different from τ.[14]

Similarly, we define the (negation of the) bias in $\hat{h}_{EM}$ as

$$b_{EM} = h - \mathbb{E}[\hat{h}_{EM}] = \delta_{EM} - \zeta_{EM}, \tag{A.3}$$

where ζ_{EM} is the model bias in $\hat{h}_{EM}$ and δ_{EM} is a constant that we subtract from $\hat{h}_{EM}$ to account for settings in which EM uses a threshold different from τ.

Typically, $\delta_{EM} = 0$ because EM is assumed to be unbiased (even though it may suffer from modeling bias). If this is the case, then we obtain the following corollary.

Corollary A.1. *If $\delta_{EM} = 0$, then $\hat{h}_{OM}$ leads to lower classification error than $\hat{h}_{EM}$ if*

$$\frac{b}{m} + \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m} < 1 - \sqrt{\gamma}, \tag{A.4}$$

where ζ_{OM} and ζ_{EM} are the model biases in $\hat{h}_{OM}$ and $\hat{h}_{EM}$, respectively.

Equation (A.4) is very similar to Equation (9) in Theorem 1, but it includes an additional term on the left side of the inequality (the second term) that represents the impact on boundary bias that results from estimating causal effects with a smaller data set.

It is well-known in the machine learning community that, when training sets are smaller, simpler models often perform better because the increase in errors caused by model bias is offset by the decrease in errors caused by variance. Thus, when causal effects are estimated using a smaller sample, we should, in general, expect the magnitude of model bias to increase as a result of the need for an algorithm that allows for less complexity in order to avoid overfitting. This, in turn, would affect the boundary bias.

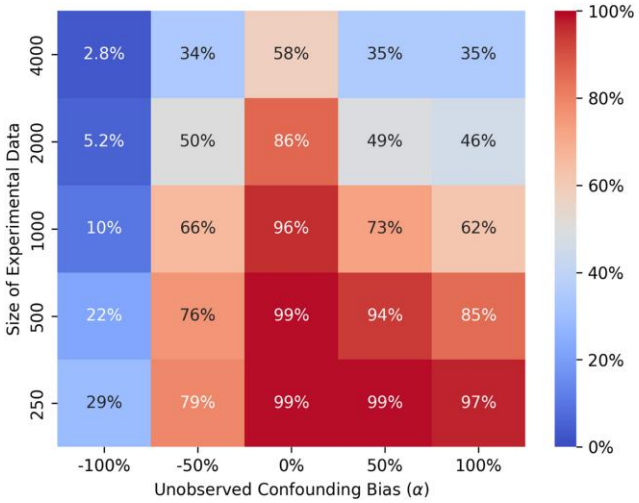
In classification tasks, however, the interplay between variance and bias implies that variance may attenuate the impact of model bias on classification performance. In the OM-versus-EM comparison (Equation (A.4)), this is reflected in the term $\zeta_{EM}\sqrt{\gamma}$: the model bias in EM is less consequential when its variance with respect to OM is large. Thus, even though we should, in general, expect EM to have more model bias than OM, model bias may comparatively affect OM more. For example, when model bias is harmful for intervention decisions ($m\zeta_{OM} < 0$ and $m\zeta_{EM} < 0$), we have the following.

Corollary A.2. *If $m\zeta_{OM} < 0$, $m\zeta_{EM} < 0$, and $\sqrt{\gamma} < \zeta_{OM}/\zeta_{EM}$, then*

$$0 < \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m}. \tag{A.5}$$

The condition $\sqrt{\gamma} < \frac{\zeta_{OM}}{\zeta_{EM}}$ is satisfied when using a larger data set to train OM reduces variance more than it decreases bias. This condition naturally holds in the straightforward scenario in which the same modeling procedure is used to train both OM and EM (so $\zeta_{OM} = \zeta_{EM}$), but EM suffers from greater variance because it was trained on a smaller sample (so $\gamma \leq 1$). Equation (A.5) suggests that OM is less likely to be preferable over EM when harmful model bias increases boundary bias, and the reduction in variance is more significant than the reduction in bias. On the other hand, the inequality in Equation (A.5) reverses if model bias is beneficial for intervention decisions ($m\zeta_i > 0, i \in \{1,2\}$); in this case, OM becomes more likely to be preferable as long as the reduction in variance outweighs the reduction in bias.

Figure B.1. (Color online) Percentage of Data Sets in Which OM Performs at Least as Well as EM with Honest Estimation



Appendix B. Experimental Analysis When Incorporating Honesty Estimation

Figure 3 contrasts OM and EM when a causal forest without honest estimation is used to estimate causal effects. In comparison, Figure B.1 in this appendix presents the results using a 50:50 split for honest estimation, which is the default option. The percentage of data sets in which OM outperforms EM is generally higher in Figure B.1 than in Figure 3. As previously mentioned, honest estimation often degrades the estimation of CATEs, particularly when the training data set is small. As a result, confounding bias has a greater potential to reduce errors caused by variance, thereby increasing the advantage of using OM.

Appendix C. Sensitivity Analysis for the Deployment of Confounded Models

We provide an example of how sensitivity analysis can be applied to assess whether a confounded model could be useful for inferring variation in causal effects when experimental data cannot be gathered.

Sensitivity analysis is a common technique used to evaluate the robustness of observational studies to confounding bias. A popular approach involves assessing the model's sensitivity to hypothetical unmeasured confounders. Typically, this analysis explores the impact on causal estimates when such confounders are included.

To illustrate, we adapt a simple instance of sensitivity analysis to demonstrate how it could be used to evaluate whether a confounded model should be deployed. Our adaptation focuses on the sensitivity of decision making rather than the magnitude of the causal effect.

Importantly, the procedure we propose here serves as a straightforward alternative for practitioners. There are many other sensitivity analysis techniques that may be more effective (e.g., Dorie et al. 2016).

In this context, a practitioner might consider deploying the potentially confounded model ($\hat{h}$) if

$$E[\mathbf{1}(\hat{h}(\boldsymbol{X}) > \hat{\tau})h(\boldsymbol{x})] > 0, \tag{C.1}$$

where $\hat{\tau}$ is the decision boundary used to make intervention decisions with $\hat{h}$. The main challenge is that we do not know h, but we could approximate it to estimate the sensitivity of $\hat{h}$ to different degrees of bias. In what follows, we approximate h as[15]

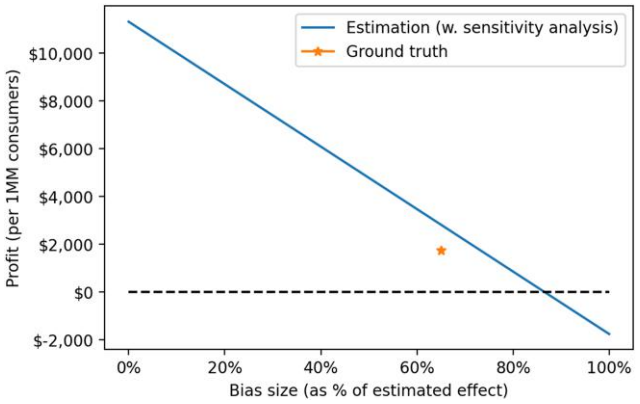
$$h(\boldsymbol{x}) \approx \hat{h}(\boldsymbol{x})(1-\alpha), \tag{C.2}$$

where α is the (unknown) degree of confounding bias (e.g., $\alpha = 0.8$ implies the actual effect is five times smaller than the confounded estimate). Note that this formulation assumes that the confounding is proportional to the size of the effect. Other formulations could be more appropriate depending on the context. Using this approximation, we can then compute how the left term in Equation (C.1) varies with respect to α to assess whether there are any plausible α values under which deploying the confounded model is worthwhile.

We showcase the proposed sensitivity analysis in the setting described in Section 6. For this example, suppose that the cost of displaying an ad is \$0.01 and the net benefit of an additional visit is \$0.49. This means the ad should be sent to an individual only if it increases the individual's likelihood of making a visit by at least 2%, so let's suppose the practitioner chooses $\hat{\tau} = 2\%$. (However, note that a larger threshold would work better if we expect the estimates to have an upward confounding bias.)

Figure C.1 demonstrates how sensitivity analysis could be used to assess whether making decisions with the confounded model would be profitable. As the analysis shows, deploying the confounded model is estimated to be profitable as long as the confounding bias does not account for more than 86% of the estimated effect (i.e., the estimate is less

Figure C.1. (Color online) Sensitivity Analysis of the Impact of Confounding Bias on Profits



than 7.14 times the size of the actual effect). In this particular case, the confounding bias accounts for 65% of the estimated effect, so it would be profitable to deploy the model.

Of course, the practitioner could still make the incorrect decision of not deploying the model if the degree of potential confounding is overestimated (i.e., the practitioner incorrectly believes that confounding could account for more than 86% of the estimated effect). Nevertheless, this example illustrates that it is not necessary to know the exact degree of potential confounding to determine whether the confounded model can be useful for decision making.

Appendix D. Proof of Analytical Derivations

Theorem D.1 follows from Theorem A.1.

Theorem D.1. *If $\zeta_{OM} = \zeta_{EM} = \delta_{EM} = 0$, then $\hat{h}_{OM}$ leads to lower classification error than $\hat{h}_{EM}$ if*

$$\frac{b}{m} < 1 - \sqrt{\gamma},$$

where $b = \delta_{OM} - \alpha$ is the uncorrectable confounding bias (the confounding bias in $\hat{h}_{OM}$ that is not corrected by using a different decision threshold for OM).

Proof. According to Equation (A.1) in Theorem A.1, $\hat{h}_{OM}$ leads to lower classification error than $\hat{h}_{EM}$ if

$$\frac{b_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma}.$$

Let $b = \delta_{OM} - \alpha$. Then, the above is

$$\frac{b - \zeta_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma}.$$

Given that $\zeta_{OM} = \zeta_{EM} = \delta_{EM} = 0$, the above is

$$\frac{b}{m} < 1 - \sqrt{\gamma}. \quad \square$$

Lemma D.1. *To simplify the notation, let $\mu = \mu(\boldsymbol{X})$ in Equation (20). Under assumptions 1 and 2, the CATE and the confounded estimand can be expressed as functions of μ and $\alpha_x^{\Delta}(\boldsymbol{X})$. The CATE is*

$$\mathbb{E}[Y^{\Delta}|\boldsymbol{X}] = \alpha_0^{\Delta} + \alpha_u^{\Delta}\mu + \alpha_x^{\Delta}(\boldsymbol{X}),$$

and the confounded estimand is

$$\mathbb{E}[Y|T=1,\boldsymbol{X}] - \mathbb{E}[Y|T=0,\boldsymbol{X}] = (\alpha_0^{\Delta} + \alpha_u^{\Delta}(\mu + \mathbb{E}[\xi|\xi > -\mu]) + \alpha_x^{\Delta}(\boldsymbol{X})) + \alpha_u^{b}(\mathbb{E}[\xi|\xi > \mu] - \mathbb{E}[\xi|\xi \leq \mu]).$$

Proof. First, we have the proof for the CATE equation:

$$\begin{aligned}\mathbb{E}[Y^{\Delta}|\boldsymbol{X}] &= \mathbb{E}_{U|X}[\mathbb{E}[Y^{\Delta}|\boldsymbol{X}, U]|\boldsymbol{X}] \\ &= \mathbb{E}[\alpha_0^{\Delta} + \alpha_u^{\Delta}U + \alpha_x^{\Delta}(\boldsymbol{X})|\boldsymbol{X}] \\ &= \alpha_0^{\Delta} + \alpha_u^{\Delta}\mu + \alpha_x^{\Delta}(\boldsymbol{X}).\end{aligned}$$

In the first line, we apply the law of total expectation. In the second line, we apply the definition of the conditional expectation of Y^{Δ}, Equation (27). We focus next on the confounded estimate:

$$
\begin{aligned}
&\mathbb{E}[Y|T=1,\boldsymbol{X}] - \mathbb{E}[Y|T=0,\boldsymbol{X}] \\
&= \mathbb{E}[Y^{\Delta} + Y^{0}|T=1,\boldsymbol{X}] - \mathbb{E}[Y^{0}|T=0,\boldsymbol{X}] \\
&= \mathbb{E}[Y^{\Delta}|T=1,\boldsymbol{X}] + (\mathbb{E}[Y^{0}|T=1,\boldsymbol{X}] - \mathbb{E}[Y^{0}|T=0,\boldsymbol{X}]) \\
&= \mathbb{E}_{U|T=1,X}[\mathbb{E}[Y^{\Delta}|T=1,\boldsymbol{X},U]|T=1,\boldsymbol{X}] + (\mathbb{E}_{U|T=1,X}[\mathbb{E}[Y^{0}|T=1,\boldsymbol{X},U]|T=1,\boldsymbol{X}] - \mathbb{E}_{U|T=0,X}[\mathbb{E}[Y^{0}|T=0,\boldsymbol{X},U]|T=0,\boldsymbol{X}]) \\
&= \mathbb{E}_{U|T=1,X}[\mathbb{E}[Y^{\Delta}|\boldsymbol{X},U]|T=1,\boldsymbol{X}] + (\mathbb{E}_{U|T=1,X}[\mathbb{E}[Y^{0}|\boldsymbol{X},U]|T=1,\boldsymbol{X}] - \mathbb{E}_{U|T=0,X}[\mathbb{E}[Y^{0}|\boldsymbol{X},U]|T=0,\boldsymbol{X}]) \\
&= \mathbb{E}[\alpha_0^{\Delta} + \alpha_u^{\Delta}U + \alpha_x^{\Delta}(\boldsymbol{X})|T=1,\boldsymbol{X}] + (\mathbb{E}[\alpha_0^{b} + \alpha_u^{b}U + \alpha_x^{b}(\boldsymbol{X})|T=1,\boldsymbol{X}] - \mathbb{E}[\alpha_0^{b} + \alpha_u^{b}U + \alpha_x^{b}(\boldsymbol{X})|T=0,\boldsymbol{X}]) \\
&= \alpha_0^{\Delta} + \alpha_x^{\Delta}(\boldsymbol{X}) + \alpha_u^{\Delta}\mathbb{E}[U|T=1,\boldsymbol{X}] + \alpha_u^{b}(\mathbb{E}[U|T=1,\boldsymbol{X}] - \mathbb{E}[U|T=0,\boldsymbol{X}]) \\
&= \alpha_0^{\Delta} + \alpha_x^{\Delta}(\boldsymbol{X}) + \alpha_u^{\Delta}\mathbb{E}[U|U>0,\boldsymbol{X}] + \alpha_u^{b}(\mathbb{E}[U|U>0,\boldsymbol{X}] - \mathbb{E}[U|U\le 0,\boldsymbol{X}]) \\
&= \alpha_0^{\Delta} + \alpha_x^{\Delta}(\boldsymbol{X}) + \alpha_u^{\Delta}\mathbb{E}[\mu+\xi|\xi>-\mu,\boldsymbol{X}] + \alpha_u^{b}(\mathbb{E}[\mu+\xi|\xi>-\mu,\boldsymbol{X}] - \mathbb{E}[\mu+\xi|\xi\le-\mu,\boldsymbol{X}]) \\
&= (\alpha_0^{\Delta} + \alpha_u^{\Delta}(\mu + \mathbb{E}[\xi|\xi>-\mu,\boldsymbol{X}]) + \alpha_x^{\Delta}(\boldsymbol{X})) + \alpha_u^{b}(\mathbb{E}[\xi|\xi>-\mu,\boldsymbol{X}] - \mathbb{E}[\xi|\xi\le-\mu,\boldsymbol{X}]) \\
&= (\alpha_0^{\Delta} + \alpha_u^{\Delta}(\mu + \mathbb{E}[\xi|\xi>-\mu]) + \alpha_x^{\Delta}(\boldsymbol{X})) + \alpha_u^{b}(\mathbb{E}[\xi|\xi>-\mu] - \mathbb{E}[\xi|\xi\le-\mu]) \\
&= (\alpha_0^{\Delta} + \alpha_u^{\Delta}(\mu + \mathbb{E}[\xi|\xi>-\mu]) + \alpha_x^{\Delta}(\boldsymbol{X})) + \alpha_u^{b}(\mathbb{E}[\xi|\xi>\mu] - \mathbb{E}[\xi|\xi\le\mu]).
\end{aligned}
$$

In the first line, we apply the definition of Y, Equation (19). In the third line, we apply the law of total expectation. In the fourth line, we apply conditional independence from T given U because T is a function of U. In the fifth line, we apply the definitions of the conditional expectations of Y^{0} and Y^{Δ} in Equations (26) and (27). In the seventh line, we apply the definition of T, Equation (18). In the eighth line, we apply the definition of U, Equation (20). In the 10th line, we apply the independence property of ξ. In the 11th line, we apply the symmetry property of ξ, Equation (23). □

Lemma D.2. *The difference between the expected treatment propensity of the treated and the untreated is*

$$\mathbb{E}[U|T=1,\boldsymbol{X}] - \mathbb{E}[U|T=0,\boldsymbol{X}] = \mathbb{E}[\xi|\xi>\mu] - \mathbb{E}[\xi|\xi\le\mu].$$

Under assumption 1*, this difference is decreasing when* $\mu<0$ *and increasing when* $\mu>0$.

Proof. First, we show the derivation of the difference of the expected treatment propensities:

$$
\begin{aligned}
&\mathbb{E}[U|T=1,\boldsymbol{X}] - \mathbb{E}[U|T=0,\boldsymbol{X}] \\
&= \mathbb{E}[U|U>0,\boldsymbol{X}] - \mathbb{E}[U|U\le 0,\boldsymbol{X}] \\
&= \mathbb{E}[\mu(\boldsymbol{X})+\xi|\xi>-\mu(\boldsymbol{X}),\boldsymbol{X}] - \mathbb{E}[\mu(\boldsymbol{X})+\xi|\xi\le-\mu(\boldsymbol{X}),\boldsymbol{X}] \\
&= \mathbb{E}[\xi|\xi>-\mu] - \mathbb{E}[\xi|\xi\le-\mu] \\
&= \mathbb{E}[\xi|\xi>\mu] - \mathbb{E}[\xi|\xi\le\mu].
\end{aligned}
$$

In the first line, we apply the definition of T, Equation (18). In the second line, we apply the definition of U, Equation (20). In the third line, we apply the independence property of ξ. In the fourth line, we apply the symmetry property of ξ, Equation (23).

We show next that this quantity increases with μ when $\mu>0$ and decreases with μ when $\mu<0$. Consider first the derivative,

$$\frac{\partial}{\partial\mu}(\mathbb{E}[\xi|\xi>\mu] - \mathbb{E}[\xi\le\mu]) = \frac{f_{\mu}}{1-F_{\mu}}(\mathbb{E}[\xi|\xi>\mu] - \mu) - \frac{f_{\mu}}{F_{\mu}}(-\mathbb{E}[\xi|\xi\le\mu] + \mu),$$

where f and F are the PDF and CDF of ξ, $f_u = f(u)$, and $F_u = F(u)$. The derivative is zero when $\mu=0$. We need to show that the derivative is positive when $\mu>0$ and negative when $\mu<0$.

Before we proceed, we first derive the value of F_{μ}. Given that $\mathbb{E}[\xi]=0$, we have that

$$\mathbb{E}[\xi] = \mathbb{P}[\xi>\mu]\mathbb{E}[\xi|\xi>\mu] + \mathbb{P}[\xi\le\mu]\mathbb{E}[\xi|\xi\le\mu] = 0.$$

To simplify the notation, let $\alpha = \mathbb{E}[\xi|\xi>\mu]$, $\beta = \mathbb{E}[\xi|\xi\le\mu]$. Then,

$$
\begin{aligned}
(1-F_{\mu})\alpha + F_{\mu}\beta &= 0 \\
(-\alpha+\beta)F_{\mu} + \alpha &= 0 \\
F_{\mu} &= \frac{\alpha}{\alpha-\beta}.
\end{aligned}
$$

Case for $\mu > 0$: First, we consider$\mu > 0$ and prove that the derivative is positive with respect to μ in this case.

$$\frac{f_\mu}{1-F_\mu}(\mathbb{E}[\xi\,|\,\xi > \mu] - \mu) - \frac{f_\mu}{F_\mu}(-\mathbb{E}[\xi\,|\,\xi \le \mu] + \mu) > 0$$

$$F_\mu(\mathbb{E}[\xi\,|\,\xi > \mu] - \mu) - (1-F_\mu)(-\mathbb{E}[\xi\,|\,\xi \le \mu] + \mu) > 0.$$

We apply the definitions of α, β, and F_μ:

$$\begin{aligned}
\frac{\alpha}{\alpha-\beta}(\alpha-\mu) - (1-\frac{\alpha}{\alpha-\beta})(\mu-\beta) &> 0\\
\alpha(\alpha-\mu) - (\alpha-\beta-\alpha)(\mu-\beta) &> 0\\
\alpha(\alpha-\mu) + \beta(\mu-\beta) &> 0\\
\alpha^2 - \alpha\mu + \beta\mu - \beta^2 &> 0\\
(\alpha+\beta)(\alpha-\beta) - (\alpha-\beta)\mu &> 0\\
(\alpha-\beta)(\alpha+\beta-\mu) &> 0\\
\alpha+\beta-\mu &> 0.
\end{aligned}$$

To prove the derivative is positive, we need to prove that $\alpha + \beta > \mu$.

F is strictly concave when $\mu > 0$, based on Equation (25). Given that the point (μ, F_μ) is between the points $(0, F_0)$ and (α, F_α), then

$$\mu = \frac{\mu}{\alpha}\cdot\alpha + \left(1-\frac{\mu}{\alpha}\right)\cdot 0,$$

so based on the property of a concave function, we can get

$$\begin{aligned}
F_\mu &> \frac{\mu}{\alpha}F_\alpha + \left(1-\frac{\mu}{\alpha}\right)F_0\\
F_\mu &> \frac{\mu}{\alpha}F_\alpha + \frac{1}{2}\left(1-\frac{\mu}{\alpha}\right)\\
2\alpha F_\mu &> 2\mu F_\alpha + \alpha - \mu\\
(2F_\mu - 1)\alpha &> (2F_\alpha - 1)\mu\\
\left(\frac{2\alpha}{\alpha-\beta} - 1\right)\alpha &> (2F_\alpha - 1)\mu\\
\frac{\alpha}{\alpha-\beta}(\alpha+\beta) &> (2F_\alpha - 1)\mu\\
F_\mu(\alpha+\beta) &> (2F_\alpha - 1)\mu\\
\alpha+\beta &> \frac{2F_\alpha - 1}{F_\mu}\mu,
\end{aligned}$$

where the second line is based on the symmetry property of ξ, Equation (23).

To ensure that the inequality $\alpha + \beta > \mu$ holds true, we need to prove that $(2F_\alpha - 1)/F_\mu > 1$.

Jensen's inequality implies that, given a random variable Z and a strictly concave function ϕ, then

$$\phi(\mathbb{E}[Z]) > \mathbb{E}[\phi(Z)].$$

Because F is strictly concave when $\mu > 0$, we have that

$$F_\alpha = F(\mathbb{E}[\xi\,|\,\xi > \mu]) > \mathbb{E}[F(\xi)\,|\,\xi > \mu].$$

The PDF of ξ conditional on $\xi > \mu$ is

$$f(\xi\,|\,\xi > \mu) = \frac{f(\xi)}{P[\xi > \mu]} = \frac{f(\xi)}{1-F_\mu}.$$

Then, we have that

$$\begin{aligned}\mathbb{E}[F(\xi)|\xi > \mu] &= \int_{\mu}^{\infty} F(\xi) f(\xi|\xi > \mu) d\xi \\ &= \int_{\mu}^{\infty} F(\xi) \frac{f(\xi)}{1 - F_{\mu}} d\xi \\ &= \int_{F_{\mu}}^{1} F(\xi) \frac{1}{1 - F_{\mu}} dF(\xi) \\ &= \frac{1}{2(1 - F_{\mu})}(1 - F_{\mu}^2) \\ &= \frac{1 + F_{\mu}}{2}.\end{aligned}$$

So we have that

$$\begin{aligned}F_{\alpha} &> \frac{1 + F_{\mu}}{2} \\ \frac{2F_{\alpha} - 1}{F_{\mu}} &> 1.\end{aligned}$$

Therefore, when $\mu > 0$,

$$\alpha + \beta > \frac{2F_{\alpha} - 1}{F_{\mu}} \mu > \mu,$$

which implies the derivative is positive when $\mu > 0$.

Case for $\mu < 0$: We show next that the derivative is negative with respect to μ when $\mu < 0$. We need to show that

$$\begin{aligned}\frac{f_{\mu}}{1 - F_{\mu}}(\mathbb{E}[\xi|\xi > \mu] - \mu) - \frac{f_{\mu}}{F_{\mu}}(-\mathbb{E}[\xi|\xi \le \mu] + \mu) < 0 \\ F_{\mu}(\mathbb{E}[\xi|\xi > \mu] - \mu) - (1 - F_{\mu})(-\mathbb{E}[\xi|\xi \le \mu] + \mu) < 0.\end{aligned}$$

We apply the definitions of α, β, and F_{μ}:

$$\begin{aligned}\frac{\alpha}{\alpha - \beta}(\alpha - \mu) - \left(1 - \frac{\alpha}{\alpha - \beta}\right)(\mu - \beta) < 0 \\ (\alpha - \beta)(\alpha + \beta - \mu) < 0 \\ \alpha + \beta - \mu < 0.\end{aligned}$$

To prove the derivative is positive, we need to prove that $\alpha + \beta < \mu$.

F is strictly convex when $\mu < 0$, based on Equation (24). Given that the point (μ, F_{μ}) is between the points $(0, F_0)$ and (β, F_{β}), then

$$\mu = \frac{\mu}{\beta} \cdot \beta + \left(1 - \frac{\mu}{\beta}\right) \cdot 0,$$

so based on the property of a convex function, we can get

$$\begin{aligned}F_{\mu} &< \frac{\mu}{\beta} F_{\beta} + \left(1 - \frac{\mu}{\beta}\right) F_0 \\ F_{\mu} &< \frac{\mu}{\beta} F_{\beta} + \frac{1}{2}\left(1 - \frac{\mu}{\beta}\right) \\ 2\beta F_{\mu} &> 2\mu F_{\beta} + \beta - \mu \\ (2F_{\mu} - 1)\beta &> (2F_{\beta} - 1)\mu \\ \left(\frac{2\alpha}{\alpha - \beta} - 1\right)\beta &> (2F_{\beta} - 1)\mu \\ \frac{\beta}{\alpha - \beta}(\alpha + \beta) &> (2F_{\beta} - 1)\mu \\ \alpha + \beta &< \frac{(2F_{\beta} - 1)(\alpha - \beta)}{\beta} \mu,\end{aligned}$$

where the second line is based on the symmetry property of ξ, Equation (23).

To ensure that the inequality $\alpha+\beta<\mu$ holds true, we need to prove that $(2F_\beta-1)(\alpha-\beta)/\beta>1$ as $\mu<0$. Jensen's inequality implies that, given a random variable Z and a strictly convex function ϕ, then

$$\phi(\mathbb{E}[Z])<\mathbb{E}[\phi(Z)].$$

Because F is strictly convex when $\mu<0$, we have that

$$F_\beta=F(\mathbb{E}[\xi\,|\,\xi\le\mu])<\mathbb{E}[F(\xi)|\xi\le\mu].$$

The PDF of ξ conditional on $\xi\le\mu$ is

$$f(\xi\,|\,\xi\le\mu)=\frac{f(\xi)}{P(\xi\le\mu)}=\frac{f(\xi)}{F_\mu}.$$

Then, we have that

$$\begin{aligned}\mathbb{E}[F(\xi)|\xi\le\mu]&=\int_{-\infty}^{\mu}F(\xi)f(\xi|\xi\le\mu)d\xi\\&=\int_{-\infty}^{\mu}F(\xi)\frac{f(\xi)}{F_\mu}d\xi\\&=\int_0^{F_\mu}F(\xi)\frac{1}{F_\mu}dF(\xi)\\&=\frac{1}{2F_\mu}(F_\mu^2-0)\\&=\frac{F_\mu}{2}.\end{aligned}$$

So we have that

$$\begin{aligned}F_\beta&<\frac{F_\mu}{2}\\F_\beta&<\frac{\alpha}{2(\alpha-\beta)}\\2F_\beta(\alpha-\beta)&<\alpha\\2F_\beta(\alpha-\beta)-(\alpha-\beta)&<\alpha-(\alpha-\beta)\\(2F_\beta-1)(\alpha-\beta)&<\beta\\\frac{(2F_\beta-1)(\alpha-\beta)}{\beta}&>1.\end{aligned}$$

Thus,

$$\alpha+\beta<\frac{(2F_\beta-1)(\alpha-\beta)}{\beta}\mu<\mu,$$

and the derivative is negative when $\mu<0$. □

Theorem D.2. *Under assumptions* 1–3, *a larger confounded estimand of the CATE also implies a larger CATE if the following inequality is true for all individuals in the population:*

$$\alpha_u^\Delta\alpha_u^b(\mathbb{P}[T=1|\boldsymbol{X}]-0.5)>0.$$

Proof. Suppose we have individuals i and j with expected treatment propensities μ_i and μ_j.

Consider first the case in which $\mu_i=\mu_j$. Let the feature vectors of the individuals be $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Lemma D.1 implies that both the CATE and the confounded estimand are functions of μ and $\alpha_x^\Delta(x)$. Therefore, if $\alpha_x^\Delta(\boldsymbol{x}_i)=\alpha_x^\Delta(\boldsymbol{x}_j)$, both individuals have the same CATE and the same confounded estimand. Without loss of generality, suppose that $\alpha_x^\Delta(\boldsymbol{x}_i)>\alpha_x^\Delta(\boldsymbol{x}_j)$. Then, Lemma D.1 implies that both the CATE and the confounded estimand are larger for individual i than for j.

We consider next the case in which the treatment propensity is greater for one of the individuals. Without loss of generality, assume that $\mu_i>\mu_j$. If $\alpha_u^\Delta>0$, then

$$\begin{aligned}&\mu_i>\mu_j\\&\Rightarrow\mu_i+\frac{\alpha_x^\Delta(\boldsymbol{x}_i)}{\alpha_u^\Delta}>\mu_j+\frac{\alpha_x^\Delta(\boldsymbol{x}_j)}{\alpha_u^\Delta}\\&\Rightarrow\alpha_u^\Delta\mu_i+\alpha_x^\Delta(\boldsymbol{x}_i)>\alpha_u^\Delta\mu_j+\alpha_x^\Delta(\boldsymbol{x}_j)\\&\Rightarrow\alpha_0^\Delta+\alpha_u^\Delta\mu_i+\alpha_x^\Delta(\boldsymbol{x}_i)>\alpha_0^\Delta+\alpha_u^\Delta\mu_j+\alpha_x^\Delta(\boldsymbol{x}_j)\\&\beta_i>\beta_j,\end{aligned}$$

where β_i and β_j are the CATEs of individuals i and j, respectively. The second line comes from assumption 3, Equation (28). The fifth line comes from Lemma D.1. In contrast, if $\alpha_u^\Delta < 0$, then $\beta_i < \beta_j$ because the inequality in the third line reverses.

Therefore, we want to prove that, if $\alpha_u^\Delta \alpha_u^b(\mathbb{P}[T=1|\boldsymbol{X}] - 0.5) > 0$, then the confounded estimand of i is greater than that of j when $\alpha_u^\Delta > 0$, and it is lower when $\alpha_u^\Delta < 0$.

Based on Lemma D.1, the confounded estimand is

$$\begin{aligned}&(\alpha_0^\Delta + \alpha_u^\Delta(\mu + \mathbb{E}[\xi|\xi > -\mu]) + \alpha_x^\Delta(\boldsymbol{X})) + \alpha_u^b(\mathbb{E}[\xi|\xi > \mu] - \mathbb{E}[\xi|\xi \le \mu])\\ &= (\alpha_0^\Delta + \alpha_u^\Delta \mathbb{E}[U|U > 0] + \alpha_x^\Delta(\boldsymbol{X})) + \alpha_u^b(\mathbb{E}[\xi|\xi > \mu] - \mathbb{E}[\xi|\xi \le \mu]),\end{aligned}$$

given the definition of U in Equation (20). A larger confounded estimand implies a larger CATE if both of these terms increase with μ when $\alpha_u^\Delta > 0$ and decrease when $\alpha_u^\Delta < 0$.

We focus first on the left term: $\alpha_0^\Delta + \alpha_u^\Delta \mathbb{E}[U|U > 0] + \alpha_x^\Delta(\boldsymbol{X})$. Let $U_i = \mu_i + \xi$ and $U_j = \mu_j + \xi$. If $\mu_i > \mu_j$:

$$\begin{aligned}&\mathbb{E}[U_i|U_i > 0]\\ &= \mathbb{E}[U_j + (\mu_i - \mu_j)|U_j > -(\mu_i - \mu_j)])\\ &> \mathbb{E}[U_j|U_j > 0].\end{aligned}$$

The above follows from the log-concavity property in assumption 1. Additionally, assumption 3 implies that α_x^Δ is nondecreasing in μ when $\alpha_u^\Delta > 0$ and nonincreasing in μ when $\alpha_u^\Delta > 0$. Therefore, the left term increases with μ when $\alpha_u^\Delta > 0$ and decreases with μ when $\alpha_u^\Delta < 0$.

Next, we focus on the right term: $\alpha_u^b(\mathbb{E}[\xi|\xi > \mu] - \mathbb{E}[\xi|\xi \le \mu])$. Let $g(\mu) = \mathbb{E}[\xi|\xi > \mu] - \mathbb{E}[\xi|\xi \le \mu]$. Lemma D.2 implies that g is increasing when $\mu > 0$ and decreasing when $\mu < 0$. Note that, because the PDF of ξ is symmetric around zero, based on Equation (23), we have $\mathbb{P}[T=1|\boldsymbol{X}] > 0.5$ when $\mu > 0$ and $\mathbb{P}[T=1|\boldsymbol{X}] < 0.5$ when $\mu < 0$. Therefore, g is increasing when $\mathbb{P}[T=1|\boldsymbol{X}] > 0.5$ and decreasing when $\mathbb{P}[T=1|\boldsymbol{X}] < 0.5$.

Based on the above, the term $\alpha_u^b g(\mu)$ is increasing when $\alpha_u^b(\mathbb{P}[T=1|\boldsymbol{X}] - 0.5) > 0$ and decreasing when $\alpha_u^b(\mathbb{P}[T=1|\boldsymbol{X}] - 0.5) < 0$. If the inequality $\alpha_u^\Delta \alpha_u^b(\mathbb{P}[T=1|\boldsymbol{X}] - 0.5) > 0$ holds, then the term is increasing when $\alpha_u^\Delta > 0$ and decreasing when $\alpha_u^\Delta < 0$.

Because both terms in the confounded estimand increase when $\alpha_u^\Delta > 0$ and decrease when $\alpha_u^\Delta < 0$, the individual with the larger confounded estimand correspond to the individual with the larger CATE. □

Theorem D.3. *Let h be the average treatment effect given a specific set of feature values, and let $\hat{h}_{OM}$ and $\hat{h}_{EM}$ be the effect estimates provided by two causal effect models. If $\hat{h}_{OM}$ and $\hat{h}_{EM}$ are normally distributed (conditional on the set of feature values), then $\hat{h}_{OM}$ leads to lower decision error than $\hat{h}_{EM}$ for individuals with those feature values if*

$$\frac{b_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma},$$

where $b_{OM} = h - \mathbb{E}[\hat{h}_{OM}]$ is the (negation of the) bias in $\hat{h}_{OM}$, $b_{EM} = h - \mathbb{E}[\hat{h}_{EM}]$ is the (negation of the) bias in $\hat{h}_{EM}$, $\gamma = \frac{Var[\hat{h}_{OM}]}{Var[\hat{h}_{EM}]}$ is the ratio of the variances of $\hat{h}_{EM}$ and $\hat{h}_{OM}$, and $m = h - \tau$ is the causal margin—the distance between the average treatment effect and the optimal decision boundary (or threshold) when h is used to make decisions.

Proof. Let $\hat{a}_1$ and $\hat{a}_2$ be the decisions made with $\hat{h}_{OM}$ and $\hat{h}_{EM}$ when τ is the decision boundary or threshold for both scoring functions:

$$\begin{aligned}\hat{a}_1 &= \mathbf{1}(\hat{h}_{OM} \ge \tau)\\ \hat{a}_2 &= \mathbf{1}(\hat{h}_{EM} \ge \tau).\end{aligned}$$

Assume that $\hat{h}_{OM}$ and $\hat{h}_{EM}$ each follow a normal distribution, and let $\omega = a^*\omega_1 + (1 - a^*)\omega_2$. Then, according to Equation (7), $\hat{h}_{OM}$ leads to lower decision error than $\hat{h}_{EM}$ if

$$\begin{aligned}&\tilde{\Phi}\left[\text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{OM}] - \tau}{\sqrt{Var[\hat{h}_{OM}]}}\right]\omega\\ &< \tilde{\Phi}\left[\text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{EM}] - \tau}{\sqrt{Var[\hat{h}_{EM}]}}\right]\omega\\ &\Rightarrow \text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{OM}] - \tau}{\sqrt{Var[\hat{h}_{OM}]}} > \text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{EM}] - \tau}{\sqrt{Var[\hat{h}_{EM}]}}.\end{aligned}$$

Let $\gamma = Var[\hat{h}_{OM}]/Var[\hat{h}_{EM}]$. Then,

$$\text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{OM}]-\tau}{\sqrt{\gamma Var[\hat{h}_{EM}]}} > \text{sign}(h-\tau)\frac{\mathbb{E}[\hat{h}_{EM}]-\tau}{\sqrt{Var[\hat{h}_{OM}]}}$$

$$\text{sign}(h-\tau)(\mathbb{E}[\hat{h}_{OM}]-\tau) > \text{sign}(h-\tau)(\mathbb{E}[\hat{h}_{EM}]-\tau)\sqrt{\gamma}.$$

Let $b_{OM} = h - \mathbb{E}[\hat{h}_{OM}]$ and $b_{EM} = h - \mathbb{E}[\hat{h}_{EM}]$. Then,

$$\text{sign}(h-\tau)(h-\tau-b_{OM}) > \text{sign}(h-\tau)(h-\tau-b_{EM})\sqrt{\gamma}.$$

Let $m = h - \tau$. Then,

$$\text{sign}(m)(m-b_{OM}) > \text{sign}(m)(m-b_{EM})\sqrt{\gamma}$$

$$\frac{m-b_{OM}}{m} > \frac{m-b_{EM}}{m}\sqrt{\gamma}$$

$$1-\frac{b_{OM}}{m} > \left(1-\frac{b_{EM}}{m}\right)\sqrt{\gamma}.$$

Therefore, $\hat{h}_{OM}$ leads to lower decision error than $\hat{h}_{EM}$ if

$$\frac{b_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma}. \quad \square$$

Corollary D.1. *If $\delta_{EM} = 0$, then $\hat{h}_{OM}$ leads to lower classification error than $\hat{h}_{EM}$ if*

$$\frac{b}{m} + \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m} < 1 - \sqrt{\gamma},$$

where ζ_{OM} and ζ_{EM} are the model biases in $\hat{h}_{OM}$ and $\hat{h}_{EM}$, respectively.

Proof. Corollary D.1 follows from Theorem D.3. According to Equation (A.1), $\hat{h}_{OM}$ leads to lower classification error than $\hat{h}_{EM}$ if

$$\frac{b_{OM}}{m} < 1 + \left(\frac{b_{EM}}{m} - 1\right)\sqrt{\gamma}.$$

Let $b = \delta_{OM} - \alpha$. Then, the above is

$$\frac{b-\zeta_{OM}}{m} < 1 + \left(\frac{\delta_{EM}-\zeta_{EM}}{m} - 1\right)\sqrt{\gamma}.$$

Given that $\delta_{EM} = 0$, the above is

$$\frac{b}{m} + \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m} < 1 - \sqrt{\gamma}, \quad \square$$

Corollary D.2. *If $m\zeta_{OM} < 0$, $m\zeta_{EM} < 0$, and $\sqrt{\gamma} < \zeta_{OM}/\zeta_{EM}$, then*

$$0 < \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m}.$$

Proof. We have that

$$\sqrt{\gamma} < \frac{\zeta_{OM}}{\zeta_{EM}}.$$

Because $m\zeta_{OM} < 0$ and $m\zeta_{EM} < 0$, then

$$\frac{\zeta_{EM}}{m}\sqrt{\gamma} > \frac{\zeta_{OM}}{m}.$$

We rearrange terms and

$$0 < \frac{\zeta_{EM}\sqrt{\gamma} - \zeta_{OM}}{m}. \quad \square$$

Endnotes

[1] This is formally defined as the difference between the expected value of the causal effect estimate and the causal effect that is the subject of the estimation.

[2] This threshold-based approach to defining optimal decision making does not consider settings in which treating the individuals with the largest effects is not always the best choice, such as when the decision maker faces complex constraints. We discuss such settings in more detail in Section 7.2.

[3] This assumption may not always hold as γ depends on the specific methods used to estimate OM and EM. For example, double/debiased machine learning methods may reduce confounding bias at the expense of substantially higher variance (Gordon et al. 2023), potentially leading to $\gamma > 1$.

[4] Data is available here: https://jenniferhill7.wixsite.com/acic-2016/competition.

[5] Here, the threshold is set to $\tau = 0$. The threshold choice affects whether confounding bias helps or harms decision making (as discussed in Section 3.3), but it does not alter the fundamental trade-off between data quantity and confounding that drives the relative performance of OM and EM. A higher threshold increases the likelihood that not treating is the optimal action, making overestimation ($\alpha > 0$) more problematic when the effect is generally positive ($\beta > 0$) and underestimation ($\alpha < 0$) less problematic. The reverse holds when the effect is generally negative ($\beta < 0$). These dynamics are a direct consequence of the theoretical framework, and although additional empirical analyses could illustrate these effects, they do not change the core insights of the paper.

[6] If this is a concern in a specific application, one could conduct additional analyses to detect systematic differences in the propensity of treatment between policies. For example, identifying features predictive of an increase in the probability of treatment when using an observational model rather than an experimental model can help pinpoint groups more likely to benefit or be harmed by one policy over another.

[7] Note that $\alpha = 0$ implies there is no unobserved confounding, but the observational data still suffers from confounding because of the nonrandom treatment assignment in the ACIC data sets. This explains the different results for OM and EM in the top center cell of Figure 3, in which the models are trained with the same data size.

[8] Whether this occurs depends critically on the decision logic implemented. If decisions are made based on a threshold on the estimated effect, as in this case, positive bias in the estimation of negative effects can lead to targeting more individuals and may result in incorrectly treating some of them. However, because all effects are overestimated equally in this case, if the decision logic instead specifies a quantile threshold (e.g., targeting the individuals with the top 10% estimated effects), the policy selects the same individuals as if the effects were estimated accurately. In this scenario, the overestimation has no negative impact.

[9] In practice, the goal is often to maximize the value (e.g., financial benefit) of the assignments rather than their causal impact. This may require us to account for the cost of the treatment and the benefit of the effect. What we propose can be adapted to take these aspects into consideration by changing the threshold in Equation (11) for deciding when to take action and incorporating benefits and costs in Equation (12).

[10] Note that the opportunity cost of investing in experimental data may be negligible if OM makes poor decisions because of confounding bias. In extreme cases in which the bias is severe (e.g., the model estimates negative effects for most people even though most benefit from the treatment), making decisions at random might be better than using OM, so the opportunity cost of running the experiment could actually be negative.

[11] See https://ailab.criteo.com/criteo-uplift-prediction-dataset/ for details and access to the data.

[12] For each experiment size, we sampled the original (unconfounded) training set 50 times. The results reported in the figure are the average of the 50 samples for each experiment size.

[13] Note the exponential data size increases on the x-axis.

[14] Equation (A.1) assumes that both models use the threshold τ that is optimal if decisions are made using h, but this is not necessarily the optimal threshold for either of the two models. Subtracting a constant δ from all estimates in a model is equivalent to using a different threshold for that model.

[15] This approximation may be overly simplistic in other settings as it does not consider other types of estimation error in $\hat{h}$ or how confounding bias may vary according to $\boldsymbol{X}$. Therefore, more sophisticated sensitivity analysis techniques could be more appropriate, depending on the circumstances.

References

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J. Marketing Res.* 55(1):80–98.

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.

Athey S, Chetty R, Imbens G (2020) Combining experimental and observational data to estimate treatment effects on long term outcomes. Preprint, submitted June 17, https://arxiv.org/abs/2006.09676.

Besbes O, Phillips R, Zeevi A (2010) Testing the validity of a demand model: An operations perspective. *Manufacturing Service Oper. Management* 12(1):162–183.

Bhattacharya D, Dupas P (2012) Inferring welfare maximizing treatment assignment under budget constraints. *J. Econometrics* 167(1):168–196.

Demirezen EM, Kumar S (2016) Optimization of recommender systems based on inventory. *Production Oper. Management* 25(4):593–608.

Devriendt F, Moldovan D, Verbeke W (2018) A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* 6(1):13–41.

Diemert E, Betlei A, Renaudin C, Amini MR (2018) A large scale benchmark for uplift modeling. *Proc. 24th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York).

Dorie V, Harada M, Carnegie NB, Hill J (2016) A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statist. Medicine* 35(20):3453–3470.

Feit EM, Berman R (2019) Test & roll: Profit-maximizing A/B tests. *Marketing Sci.* 38(6):1038–1058.

Feng Q, Luo S, Zhang D (2014) Dynamic inventory–pricing control under backorder: Demand estimation and policy optimization. *Manufacturing Service Oper. Management* 16(1):149–160.
Fernández-Loría C, Provost F (2020) Combining observational and experimental data to improve large-scale decision-making. *Proc. Internat. Conf. Inform. Systems* (Association for Information Systems, Atlanta), 1583.
Fernández-Loría C, Provost F (2022a) Causal classification: Treatment effect estimation vs. outcome prediction. *J. Machine Learn. Res.* 23(59):1–35.
Fernández-Loría C, Provost F (2022b) Causal decision making and causal effect estimation are not the same … and why it matters. *INFORMS J. Data Sci.* 1(1):4–16.
Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper Management* 18(1):69–88.
Frey LJ, Fisher DH (1999) Modeling decision tree performance with the power law. *Proc. Seventh Internat. Workshop Artificial Intelligence Statist.*, Proceedings of Machine Learning Research (PMLR, New York).
Friedman JH (1997) On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Mining Knowledge Discovery* 1(1):55–77.
Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. *Neural Comput.* 4(1):1–58.
Gordon BR, Moakler R, Zettelmeyer F (2023) Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Sci.* 42(4):768–793.
Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Sci.* 38(2):193–225.
Hill JL (2011) Bayesian nonparametric modeling for causal inference. *J. Comput. Graphical Statist.* 20(1):217–240.
Hirano K, Porter JR (2009) Asymptotics for statistical treatment rules. *Econometrica* 77(5):1683–1701.
Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist.* 7(1):443–470.
Kallus N, Puli AM, Shalit U (2018) Removing hidden confounding by experimental grounding. *Proc. 32nd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 10911–10920.
Kane K, Lo VS, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *J. Marketing Anal.* 2(4):218–238.
Kent DM, Paulus JK, Van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JP, et al. (2020) The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann. Internal Medicine* 172(1):35–45.
Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.
Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009) Controlled experiments on the web: Survey and practical guide. *Data Mining Knowledge Discovery* 18(1):140–181.
Manski CF (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–1246.
McFowland E III, Gangarapu S, Bapna R, Sun T (2021) A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects. *MIS Quart.* 45(4):1807–1832.
Morucci M, Noor-E-Alam M, Rudin C (2022) A robust approach to quantifying uncertainty in matching problems of causal inference. *INFORMS J. Data Sci.* 1(2):156–171.
Pearl J (2009) *Causality: Models, Reasoning and Inference* (Cambridge University Press, Cambridge, UK).
Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: A learning-curve analysis. *J. Machine Learn. Res.* 4(June):211–255.
Peysakhovich A, Lada A (2016) Combining observational and experimental data to find heterogeneous treatment effects. Preprint, submitted November 8, https://arxiv.org/abs/1611.02385.
Provost F, Fawcett T (2013) *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* (O'Reilly Media, Sebastopol, CA).
Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions.*
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
Rosenman ET, Basse G, Owen AB, Baiocchi M (2023) Combining observational and experimental datasets using shrinkage estimators. *Biometrics* 79(4):2961–2973.
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* 66(5):688–701.
Simester D, Timoshenko A, Zoumpoulis SI (2020) Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Sci.* 66(8):3412–3424.
Train KE (2009) *Discrete Choice Methods with Simulation* (Cambridge University Press, Cambridge, UK).
Verbeke W, Olaya D, Guerry MA, Van Belle J (2023) To do or not to do? Cost-sensitive causal classification with individual treatment effect estimates. *Eur. J. Oper. Res.* 305(2):838–852.
Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113(523):1228–1242.
Wooldridge JM (2015) *Introductory Econometrics: A Modern Approach*, 6th ed. (Cengage Learning, Boston).
Yahav I, Shmueli G, Mani D (2016) A tree-based approach for addressing self-selection in impact studies with big data. *MIS Quart.* 40(4):819–848.
Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E (2012) Estimating optimal treatment regimes from a classification perspective. *Stat* 1(1):103–114.
Zhao Y, Zeng D, Rush AJ, Kosorok MR (2012) Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* 107(499):1106–1118.