# Causal Decision Making and Causal Effect Estimation Are Not the Same…and Why It Matters

Carlos Fernández-Loría, Foster Provost

# Causal Decision Making and Causal Effect Estimation Are Not the Same … and Why It Matters

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

[a] Department of Information Systems, Business Statistics, and Operations Management, HKUST Business School, Hong Kong University of Science and Technology, New Territories, Hong Kong; [b] Department of Technology, Operations, and Statistics, NYU Stern School of Business, New York University, New York, New York 10012; [c] Compass Inc., New York, New York 10011
*Corresponding author
**Contact:** imcarlos@ust.hk, https://orcid.org/0000-0003-4509-3768 (CF-L); fprovost@stern.nyu.edu (FP)

**Abstract.** Causal decision making (CDM) at scale has become a routine part of business, and increasingly, CDM is based on statistical models and machine learning algorithms. Businesses algorithmically target offers, incentives, and recommendations to affect consumer behavior. Recently, we have seen an acceleration of research related to CDM and causal effect estimation (CEE) using machine-learned models. This article highlights an important perspective: CDM is not the same as CEE, and counterintuitively, accurate CEE is not necessary for accurate CDM. Our experience is that this is not well understood by practitioners or most researchers. Technically, the estimand of interest is different, and this has important implications both for modeling and for the use of statistical models for CDM. We draw on recent research to highlight three implications. (1) We should carefully consider the objective function of the causal machine learning, and if possible, optimize for accurate "treatment assignment" rather than for accurate effect-size estimation. (2) Confounding affects CDM and CEE differently. The upshot here is that for supporting CDM it may be just as good or even better to learn with confounded data as with unconfounded data. (3) Causal statistical modeling may not be necessary at all to support CDM because a proxy target for statistical modeling might do as well or better. This third observation helps to explain at least one broad common CDM practice that seems "wrong" at first blush—the widespread use of noncausal models for targeting interventions. The last two implications are particularly important in practice, as acquiring (unconfounded) data on both "sides" of the counterfactual for modeling can be quite costly and often impracticable. These observations open substantial research ground. We hope to facilitate research in this area by pointing to related articles from multiple contributing fields, most of them written in the last five years.

## 1. Introduction

Causal decision making at scale has become a routine part of business. Firms build statistical models using machine learning algorithms to support decisions about targeting all manner of interventions. The most common are offers, advertisements, retention incentives, and recommendations made to customers and prospective customers. The goal of these interventions is to affect consumer behavior, for example, to increase purchasing or to keep a customer from churning. When we talk about causal decision making in this article, we specifically mean deciding whether to "treat" a particular individual with a particular intervention.

Making explicit the causal nature of such decision making has lagged its implementation in practice. Today, data-savvy firms routinely conduct causal *evaluations* of their targeting decisions, for example, conducting A/B

tests when making offers or targeting advertisements. However, very few firms are actually building causal models to support their causal decision making.

The academic community has, of course, taken on (conditional) causal estimation, and over the last four to five years, we have seen an acceleration of research on the topic.[1] Much of this work has focused on machine learning methods for individual causal effect estimation (Dorie et al. 2019). The main motivation behind several of these methods is causal decision making (Athey and Imbens 2017, 2019; McFowland et al. 2022).

This article highlights an important perspective: causal decision making (CDM) is not the same as causal effect estimation (CEE). Counterintuitively, accurate CEE is not necessary for accurately assigning treatments to maximize causal effect. Our experience is that this is not well understood by practitioners nor by most data

science researchers—including some who routinely build models for estimating causal effects. Very recent research has started to examine this question, but much fertile ground for new research remains due to the limited attention this perspective has received. We will present three main implications for research and practice to highlight how this perspective can be enlightening and impactful, but many other—possibly even more important—implications might be revealed by future research.

The fundamental problem in many predictive tasks is often assessing whether an intervention will have an effect. For example, when attempting to stop customers from leaving, the goal is not to target the customers most likely to leave, but instead to target the customers whom the incentive will cause to stay. For many advertising settings, the advertiser's goal is not simply to target ads to people who will purchase after seeing the ad, but to target people the ad will cause to purchase. Generally, instead of simply predicting the likelihood of a positive outcome, for tasks such as these we ideally would like to predict whether a particular individual's outcome can be improved by means of an intervention or treatment (i.e., the ad or the retention incentive in the examples).

This type of causal decision making is an instance of the problem of treatment assignment (Manski 2004). Each possible course of action corresponds to a different "treatment" (such as target versus not target), and ideally, each individual is assigned to the treatment associated with the most beneficial outcome (e.g., the one with the highest profit). Treatment assignment policies may be estimated from data using statistical modeling, allowing decision makers to map individuals to the best treatment according to their characteristics (e.g., preferences, behaviors, history with the firm). There are many different ways in which one could proceed, and various methods for the estimation of treatment assignment policies from data have been proposed across different fields, including econometrics (Bhattacharya and Dupas 2012), data mining (Radcliffe and Surry 2011), and machine learning in the contextual bandit setting (Li et al. 2010).

One approach for causal decision making that is becoming popular in the research literature is to use machine-learned models to estimate causal (treatment) effects at the individual level, and then use the models to make intervention decisions automatically (Olaya et al. 2020). For instance, in cases where there is a limited advertising budget, causal effect models may be used to predict the expected effect of a promotion on each potential customer and subsequently target those individuals with the largest predicted effect.

However, causal decision making (treatment assignment) and causal effect estimation are not the same. Going back to the advertising example, suppose a firm is willing to send an offer to customers for whom it increases the probability of purchasing by at least 1%. In this case, precisely estimating individual causal effects is desirable but not necessary; the only concern is identifying those individuals for whom the effect is greater than 1%. Importantly, overestimating (underestimating) the effect has no bearing on decision making when the focal individuals have an effect greater (smaller) than 1%. This is very similar to the distinction in noncausal predictive modeling between regression or probability estimation and classification, and we will take advantage of that analogy as we proceed.

The main distinction between causal decision making and causal effect estimation is their estimand of interest. The main object of study in (heterogeneous) causal effect estimation is understanding treatment effects in heterogeneous subpopulations. On the other hand, causal effect estimates are important for causal decision making only to the extent that they help to discriminate individuals according to preferred treatments.[2]

This paper draws on recent research to highlight three implications of this distinction.

1. We should carefully consider the objective function of the causal machine learning algorithm, and if possible, it should optimize for accurate treatment assignment rather than for accurate effect estimation. A CDM practitioner reaching for a "causal" machine learning algorithm geared toward accurately estimating conditional causal effects (as most are) may be making a mistake.

2. Confounding does not have the same effect on CDM as it does on CEE. When supporting CDM, it may be as good or even better to learn with confounded data—such as data suffering from uncorrectable selection bias. Acquiring (unconfounded) data on both "sides" of the counterfactual for modeling can be quite costly and often impracticable. Furthermore, even when we have invested in unconfounded data—for example, via a controlled experiment—often we have much more confounded data, so there is a causal bias/variance tradeoff to be considered.

3. Causal statistical modeling may not be necessary at all to support CDM, because there may be (and perhaps often is) a proxy target for statistical modeling that can do as well or better. This third implication helps to explain at least one broad common CDM practice that seems "wrong" at first blush—the widespread use of noncausal models for targeting interventions.

Overall, this paper develops the perspective that what might traditionally be considered good estimates of causal effects are not necessary to make good causal decisions. The three implications above are quite important in practice, because acquiring data to estimate causal effects accurately is often complicated and expensive. Empirically, we see that results can be considerably better when modeling intervention decisions rather than causal effects. On the research side, as mentioned previously, taking this perspective reveals substantial room for novel research.

## 2. Distinguishing Between Effect Estimation and Decision Making

We use the potential outcomes framework (Rubin 1974) to define causal effects. The framework defines interventions in terms of treatment alternatives (treatment "levels"), which may range from two (e.g., to represent absence or presence of a treatment) to many more (e.g., to represent treatment doses). The framework also assumes the existence of one potential outcome associated with each treatment alternative for each individual, which represents the value that the outcome of interest would take if the individual were to be exposed to the associated treatment. Causal effects are defined in terms of the difference between two potential outcomes. Our subsequent discussions and examples focus on cases with only two intervention alternatives, but most of the ideas and insights also apply to cases with more.

The main difference between causal effect estimation and estimating the best treatment assignment is the estimand of interest. Formally, let $Y(i) \in \mathbb{R}$ be the potential outcome associated with treatment level $i$. The estimand in (heterogeneous) causal effect estimation (CEE) is

$$\tau(x) = E[Y(1) - Y(0) \mid X = x], \qquad (1)$$

which corresponds to the conditional average treatment effect (CATE) given the feature vector $X$ used to characterize individuals. Several studies have proposed the use of machine learning methods for CATE estimation (see Dorie et al. 2019 for a survey). Popular methods include Bayesian additive regression trees (Hill 2011), causal random forests (Wager and Athey 2018), and regularized causal support vector machines (Imai and Ratkovic 2013).

The counterpart in causal decision making (CDM), or treatment assignment, is

$$a^*(x) = \mathbf{1}(\tau(x) > 0), \qquad (2)$$

which corresponds to the treatment level (or action) that maximizes the expected value of the outcome of interest given the feature vector $X$.[3] Why are causal effect models so often considered the natural choice for causal decision making? Because Equation (2) is defined in terms of Equation (1): individuals should be assigned to treatment level $a^* = 1$ when their CATE is positive and to treatment level $a^* = 0$ otherwise.

However, models provide only *estimates* of causal effects, and estimates incorporate errors. It turns out that the sorts of errors incorporated by the estimation procedures are critical. Moreover, as we discuss in detail, accurate causal effect estimation is not necessary to make good intervention decisions, and more accurate estimates of causal effects may not imply better decisions.

A second important distinction between causal effect estimation and causal decision making is that the outcome of interest may differ between the two. For example, in advertising settings, machine learning could be used to predict treatment effects on purchases with the intent of making treatment decisions that maximize profits. However, the individuals for whom treatments are most effective—in terms of increasing the likelihood of purchase—are not necessarily the ones for whom the promotions are most profitable (Lemmens and Gupta 2020, Miller and Hosanagar 2020), so making decisions based on CATEs may lead to suboptimal treatment assignments in these cases. On the other hand, when the outcomes of interest in the two tasks match (e.g., when CATEs are defined in terms of profits rather than purchases), then making decisions based on the true CATEs as defined in Equation (2) leads to optimal treatment assignments. The problem with estimation errors remains.

A third distinction is that CDM may involve additional constraints that are critical for the optimization problem at hand. For example, targeting the individuals with the largest CATEs may not be optimal when there are budget constraints and costs vary across individuals. One way to proceed in such cases is to use machine learning to predict CATEs and then use those predictions to solve a downstream optimization problem that incorporates the business constraints (McFowland et al. 2022). However, decisions can improve when both the business objective and the business constraints are incorporated as part of the machine learning (Elmachtoub et al. 2020, Elmachtoub and Grigas 2021).

More generally, this work contributes to the burgeoning stream of research that recognizes that model learning is often simply a means to a higher-level goal. In our case, learning a CEE model is often only a step toward CDM, and as we demonstrate next, sometimes circumventing this step and directly optimizing for decision making can lead to better results. Similarly, another common goal when learning CEE models is to characterize and identify subpopulations affected by the treatment (McFowland et al. 2018). Once again, the end goal for this application is subtly different from CEE or CDM because the estimand of interest corresponds to the identification of the feature values of subpopulations affected by the treatment, and results can be substantially better when algorithms are designed with that specific purpose in mind (McFowland et al. 2018). In the context of marketing (Lemmens and Gupta 2020) and operations (Elmachtoub and Grigas 2021), researchers have also reported substantially better results when models are directly optimized to minimize decision-making errors instead of prediction errors. Thus, we hope this paper will help emphasize the importance of accounting for how model predictions will be used when building data-driven solutions.

Another task closely related to CDM as defined in this paper is the contextual multiarmed bandit (MAB) problem (Slivkins 2019), which consists of

estimating the arm (treatment level) with the largest reward (potential outcome) given some context (feature vector). However, an important distinction with respect to our setting is that the goal in bandit problems is to learn a CDM model while actively making treatment assignment decisions for incoming subjects. Therefore, an exploration/exploitation dilemma plays an important role in the decision-making procedure, whereas in our case, the decision maker cannot re-estimate the treatment assignment policy after making each decision.[4] Thus, our setting is commonly referred to as "offline learning" by the MAB community (Beygelzimer and Langford 2009).

Nonetheless, we hope several of the ideas discussed here will also be revealing to researchers in this community. Reward regression is a common solution for deciding which arm to choose. However, the perspective of this paper is that reward regression and estimating the arm with the largest reward are not the same task. Therefore, the implications discussed here could also help inform future methodological developments in the MAB community.

## 3. Why It Matters

Let's consider three important practical implications that result from causal decision making and causal effect estimation having different estimands of interest.

### 3.1. Choosing the Objective Function for Learning

The first implication is that a machine learning model optimized for treatment assignment can lead to better decisions than a machine learning model optimized for causal effect estimation (Fernández-Loría et al. 2020). Machine learning procedures that estimate causal effects are designed to segment individuals according to how their effects vary, but doing so is a more general problem than segmenting individuals according to how they should be treated. By focusing on a specific task (treatment assignment) rather than on a more general task (estimating causal effects), machine learning procedures can better exploit the statistical power available.

Formally, the (sometimes unstated) goal of machine learning methods that estimate causal effects is to optimize the (possibly regularized) mean squared error (MSE) for treatment effects:

$$\text{MSE}(\hat{\tau}) = E[((Y(1) - Y(0)) - \hat{\tau}(X))^2], \quad (3)$$

which is minimized when $\hat{\tau} = \tau$ (see Equation (1)). The main challenge in causal inference is that only one potential outcome is observed for any given individual, so Equation (3) cannot be calculated directly because $Y(1) - Y(0)$ is not observable. Fortunately, alternative formulations may be used to estimate the

treatment effect MSE from data (Schuler et al. 2018), allowing one to compare (and optimize) models based on how good they are at predicting causal effects.

However, optimizing causal effect predictions (by minimizing Equation (3)) is not the same as optimizing treatment assignments. We illustrate this in Figure 1, which compares the causal effect predictions made by two different models for a single individual, one with a high effect-prediction error (model 1) and another with a low effect-prediction error (model 2). The blue (dark) diamonds correspond to the "true" CATE for the individual (they are the same for both models), whereas the red dots correspond to the predictions made by the models. A larger distance between the blue diamonds and the red dots (represented by dashed lines) implies that the model has a larger MSE and therefore makes worse causal effect predictions.

In this example, the true CATE is greater than zero (and hence above the decision boundary), which implies that this individual should be treated. Therefore, if $\hat{\tau}(x) > 0$, $\hat{\tau}$ leads to an optimal treatment assignment. Consider the zero line to be the causal effect of the "do not treat" decision.[5] Figure 1 shows that the model with the larger prediction error makes the better treatment assignment because the rank ordering of the treatment alternatives (treat and not treat) when using the predicted effect is the same as the rank ordering when using the CATE. The second model makes a worse treatment assignment, even though its effect-prediction errors are smaller because the ordering is inverted. Therefore, choosing the model with the lower (and thus better) MSE leads to a worse treatment assignment decision.

**Figure 1.** (Color online) Comparison of the Causal Effect Predictions Made by Two Models
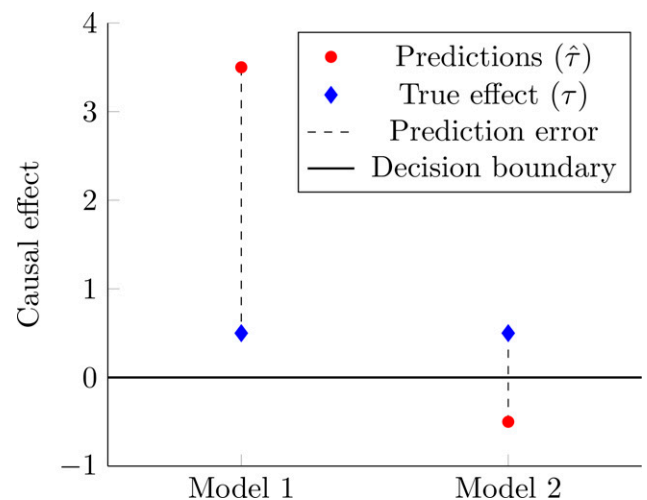


*Note.* Model 1 has a larger prediction error than model 2 but a smaller decision-making error because it predicts the sign of the causal effect correctly.

The previous discussion implies that, when evaluating models for treatment assignment, what matters is their ability to minimize the expected difference between the best potential outcome that could be obtained and the potential outcome that would be obtained by using the model. This evaluation measure is also known as expected regret (or simply "regret") in decision theory:

$$\text{Regret}(\hat{a}) = E[Y(a^*(X)) - Y(\hat{a}(X))], \qquad (4)$$

where $\hat{a}$ corresponds to the evaluated treatment assignment policy, such as $\hat{a} = \mathbf{1}(\hat{\tau}(x) > 0)$. Regret is minimized when $\hat{a} = a^*$ (see Equation (2)). As with causal effect estimation, an important challenge when learning treatment assignment policies from data is that only one potential outcome is observed for each individual. However, alternative formulations of Equation (4) allow the use of historical data to evaluate models for individual treatment assignment on an aggregate basis, based on observed treatment assignments (Li et al. 2010, Schuler et al. 2018).

Therefore, how can machine learning methods optimize for treatment assignment itself instead of optimizing for accurate effect estimation? One approach to minimizing Equation (4) using machine learning is to reduce treatment assignment to importance weighted classification (Zadrozny 2003, Zhao et al. 2012). As Zadrozny (2003) first noted, treatment assignment can be framed as a weighted classification problem where the optimal treatment assignment is the target variable (as defined in Equation (2)), and the outcomes observed under each treatment level serve to weight observations. This framing allows the use of any supervised classification algorithm to optimize for treatment assignment (Beygelzimer and Langford 2009), but methods specifically designed for solving treatment assignment as a weighted classification problem also exist (Zhao et al. 2012).

Notably, because minimizing regret is equivalent to minimizing the weighted misclassification rate as specified by Zadrozny (2003), a useful analogy in this context is to think about treatment assignment as a classification problem and to think about causal effect estimation as a probability estimation (or regression) problem; the two tasks are related but not the same.

It is well known in the predictive modeling literature that models with good classification performance are not necessarily good at estimating class probabilities and vice versa (Friedman 1997). Similarly, the bias and variance components of the estimation error in causal effect predictions may combine to influence treatment assignment in a very different way than the squared error of the predictions themselves (Fernández-Loría and Provost 2019a). Therefore, models that are relatively bad at causal effect prediction may be

good at making treatment assignments, and models that are good at causal effect prediction may be relatively bad at making treatment assignments.

Of course, in theory, the gap in performance between models optimized for causal effect estimation and models optimized for treatment assignment should disappear with unlimited data because good causal effect models should converge to optimal decision making (as shown in Equation (2)). However, empirical evidence shows that the gap can persist even with training data consisting of more than half a billion observations (Fernández-Loría et al. 2020).[6] Therefore, modeling optimal treatment assignment (rather than causal effects) can result in better causal decision making even in cases where there is plenty of data to estimate causal effects.

### 3.2. Choosing the Training Data
The second implication is that data that provide little utility for accurate causal effect estimation, due to confounding, may still be quite useful for causal decision making (Fernández-Loría and Provost 2019b). One of the main concerns when estimating causal effects from data is the possibility of bias because of confounding, such as, selection bias (Yahav et al. 2015). Formally, if $T$ represents the treatment assignment, confounding occurs whenever $E[Y(i) \mid T = i, X] \neq E[Y(i) \mid X]$, which implies that individuals who were exposed to the treatment level $i$ are systematically different from the individuals in the overall population and, in particular, different with respect to the outcome of interest. Estimating causal effects from data requires that the individuals who received the treatment (the treated) and the individuals who did not receive the treatment (the untreated) be comparable in all aspects related to the outcome, except for the assigned treatment. This assumption is known as ignorability (Rosenbaum and Rubin 1983), the backdoor criterion (Pearl 2009), and exogeneity (Wooldridge 2015). If, for example, the treated would have a better outcome than the untreated even without the treatment, then the estimation will suffer from an upward "causal"[7] bias; the treatment will be falsely attributed a positive causal effect that should be attributed to another unobserved, systematic difference between treated and untreated.

Bias caused by (unobserved) confounding is often regarded as the kiss of death in the causal inference world. Because it cannot be quantified from observational data alone, no additional amount of data can help if the goal is to estimate causal effects. In fact, with more data, the estimate will converge to a biased (and therefore arguably wrong) estimate of the effect.

Nevertheless, confounding bias does not necessarily hurt decision making (Fernández-Loría and Provost

2019b). For example, suppose that the large prediction error made by model 1 in Figure 1 is the result of using confounded training data to estimate the model. In this case, the confounding bias is clearly not hurtful for decision making. In addition, by lessening potential errors due to variance, more data can potentially correct the detrimental effect of confounding bias on decision making and even confer a beneficial effect, as we illustrate later.

The implication that avoiding confounding is less critical for CDM is particularly important 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 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 would almost certainly be biased because the people targeted with interventions are those who had been estimated to be likely to have a positive result (even without an intervention).

Some recent developments in the estimation of treatment assignment policies from observational data are specifically motivated to deal with confounding bias (Athey and Wager 2021). However, these approaches assume that the selection is observable (i.e., all confounding is captured by the feature vector $X$) or that treatment effects can be consistently estimated in some alternative way (e.g., using instrumental variables). Yet evidence from large-scale experiments suggests that confounding bias is likely to persist after controlling for observable variables (Gordon et al. 2019), and as pointed out above, correcting for confounding bias in causal effect estimates may actually lead to worse treatment assignments by increasing errors due to variance.

Another alternative is to collect data for model training using randomized experiments. Estimating causal effects from observational data often is downplayed in favor of randomized experiments due to the latter's ability to control for all confounders; the randomization procedure ensures that, on average, the treated and untreated will be comparable in all aspects other than the treatment assignment. Therefore, one could think about experiments as the natural approach to avoid confounding.

On the other hand, we should note that by our understanding of current practice (not academic research), most randomized experiments are not implemented to gather data for training models, but rather to evaluate specific hypotheses. This evaluation may include comparing different machine learning models, for example, to determine if a new model is statistically significantly better than the currently deployed model. The important point here is that the amount of data needed to *evaluate* model performance is usually much smaller than the amount of data needed to *train* accurate models of conditional treatment effects.

This is critical because randomized experiments producing enough data to train accurate CATE models may be very expensive—in many situations prohibitively expensive—forcing us to proceed using whatever data we may have at hand. They are expensive because the treatments may be costly (especially treating those for whom the treatment would have no beneficial effect) and because of the opportunity costs incurred by withholding treatment from those for whom it would be beneficial. Moreover, randomized experiments are not always possible. For example, one cannot randomly force workers to take or switch jobs to assess the impact of different jobs on future career opportunities. Yet, freelancing platforms like Upwork may be interested in making job recommendations. Business-to-business firms often cannot simply withhold services from their customers, as a consumer-oriented business might.[8]

Let's illustrate the feasibility of using confounded data for causal decision making with another example, shown in Figure 2. In this example, we need to decide whether to intervene on a particular individual. Figure 2 shows the true causal effect $\tau$ and the optimal threshold $\delta$ for deciding whether to intervene (i.e., $\delta$ considers the costs of intervening versus not intervening). In our example, $\tau > \delta$, meaning that the optimal decision for this individual is to intervene.

Of course, we do not know the true effect $\tau$, so we consider two models to estimate it: a biased model (BM) and an unbiased model (UM). Each of the panels in Figure 2 shows the sampling distributions of the causal effect estimates provided by the two models under different degrees of causal bias. One model (BM, in blue) is biased because of confounding that varies across panels—this bias can be seen in the figures, as the distributions of the BM's estimate are not centered on $\tau$. At the same time, the BM benefits from smaller variance because there are plenty of naturally occurring (biased) training data. The second model (UM, in red) is the same across all three panels, and it is unbiased—we can see that its distribution is indeed centered on $\tau$. However, the UM suffers from larger variance because it was learned using an unconfounded training data set that is smaller because of experimentation costs.

When using a model for decision making, we obtain an estimate drawn from the model's sampling distribution, and we intervene if the estimate of the causal effect is greater than (to the right of) $\delta$. Therefore, for each sampling distribution, the area to the left side of $\delta$ represents the probability of making the wrong decision. In this example, the UM usually leads to the correct decision (i.e., most of the red distribution falls to

**Figure 2.** (Color online) Sampling Distributions of the Causal Effect Estimates of a Biased Model (BM) and an Unbiased Model (UM)
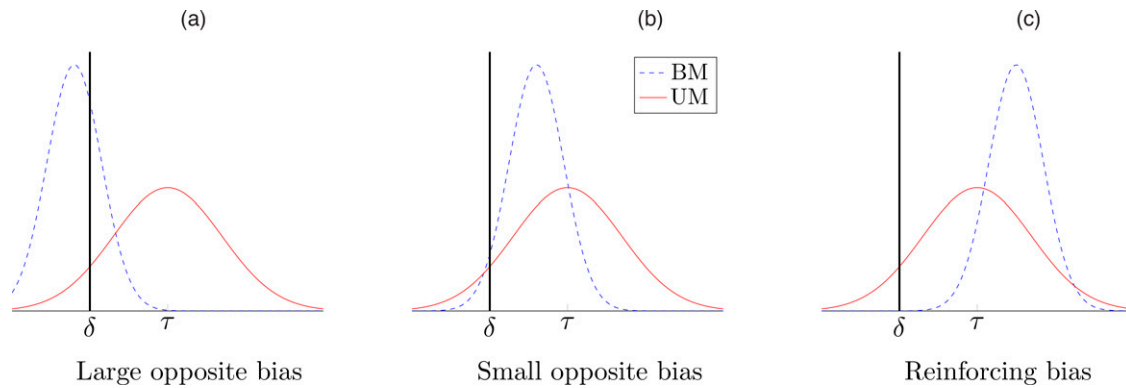


*Notes.* In this example (see details in the text), the correct decision is made when a model's estimate of $\tau$ is larger than $\delta$. For decision making, a biased model (BM) can be either worse (a), about the same (b), or even better (c) than an unbiased model (UM).

the right of $\delta$).[9] In contrast, decision making using the BM is strongly affected by the causal bias. Figure 2(a) shows a case in which more than half of the BM distribution falls to the left of $\delta$ because of a large "downward" bias that leads the BM to systematically underestimate $\tau$. Under these circumstances, the bias is so strong that the BM makes the wrong decision on average. Thus, here the UM is a better alternative, despite having larger variance.

Now consider Figure 2(b). Here the causal bias still leads the BM to systematically underestimate $\tau$, but slightly less. This small change in effect estimation bias has a large effect on decision-making accuracy; for decision making, the BM's smaller variance compensates for its bias. Whereas in Figure 2(a), the UM was a much better choice, in Figure 2(b), it is no longer clear which model is a better alternative—the two models seem equally likely to make the correct decision. It is easy to see how the BM could be preferable despite its bias if it had even smaller variance. Importantly, in this case, both the BM and the UM would converge to the optimal decision with unlimited data because their distributions are centered on the "correct side" of the threshold, implying that the detrimental effect of confounding on decision making can be corrected with larger data. Therefore, although the BM cannot recover the true causal effect, it can in fact identify the optimal decision.

Finally, Figure 2(c) shows an example in which causal bias could help. In our example, this would happen when the bias leads to an overestimation of the causal effect,[10] which lessens errors because of variance. In practice, this could occur when confounding is stronger for individuals with large effects—for example, if confounding bias is stronger for "likely buyers," but the effect of ads is also stronger for them. In such scenarios, the BM would be a better alternative than the UM for making intervention decisions, even if its causal effect estimates are highly inaccurate.

The key insight here is that optimizing to make the correct decision generally involves understanding whether a causal effect is above or below a given threshold, which is different from optimizing to reduce the magnitude of the bias in a causal effect estimate. Hence, correcting for confounding bias when learning treatment assignment policies (as proposed by Athey and Wager 2021) may hurt decision making. Furthermore, models trained with confounded data may lead to decisions that are as good (or better) than the decisions made with models trained with costly experimental data (Fernández-Loría and Provost 2019b), in particular when larger causal effects are more likely to be overestimated or when the variance reduction benefits of more and cheaper data outweigh the detrimental effect of confounding.

Although most of our discussion focuses on selection in treatment assignment as a source of confounding bias, this insight is agnostic to the source of the bias and can be generalized to other situations where the data may be problematic for causal effect estimation. For instance, the bias could be the result of violating the stable unit treatment value assumption (SUTVA), noncompliance with the treatment assignment, flaws in the experimental design, or differences between the setup in the training data and the setting where the trained model is expected to be used. The main takeaway is that data issues that make it impossible to estimate causal effects accurately do not necessarily keep us from using the data to make accurate intervention decisions.

### 3.3. Leveraging Alternative Target Variables

The third implication that arises from distinguishing between causal decision making and causal effect estimation is that, for causal decision making, modeling with alternative (potentially noncausal) target variables may work better than predicting the actual causal effects on the outcome of interest (Fernández-Loría

and Provost 2019a). To elaborate on this implication, consider again the binary treatment setting where we would like to treat those individuals for whom the treatment will have a positive effect. This elaboration defines a causal classification problem: classify individuals into treat or do-not-treat.[11]

Viewing CDM as a classification task allows us to draw an analogy with the distinctions between classification versus regression and classification versus probability estimation that have been studied in depth in noncausal predictive modeling. Specifically, many classification decisions essentially involve comparing an instance's estimated class membership probability to a threshold and then taking action based on whether the estimate is above or below the threshold.[12] In such situations, causal effect estimates are relevant for decision making only to the extent that they help to discriminate between the desired actions, for example, those who would benefit from being treated vs. those who would not.

What does this have to do with alternative target variables? For traditional classification, we know that often we can build more effective predictive models by using a proxy target variable, rather than a target variable that represents the true outcome of interest (Provost and Fawcett 2013). Proxy modeling is often applied when data on the true objective are in short supply (or even completely nonexistent). For example, in online advertising, often few data are available on the ultimate desired conversion (e.g., purchases) because of low conversion rates or cold start periods. This can pose a challenge for accurate campaign optimization. In this context, Dalessandro et al. (2015) demonstrate the effectiveness of proxy modeling and find that predictive models built based on brand site visits (which are much more common than purchases) do a remarkably good job of predicting which online consumers will purchase. Causal modeling also often suffers from scant training data; similar to noncausal predictive modeling, estimating the best action based on a proxy target may work better than directly estimating causal effects on the outcome of interest.

The use of proxy outcomes to build predictive conditional causal models should not be a major paradigm shift for researchers, because proxy or "surrogate" outcomes already are sometimes used to evaluate treatment effects (Prentice 1989, VanderWeele 2013). For example, in clinical trials, the goal is often to study the efficacy of an intervention on outcomes such as patients' long-term health or survival rate. However, the primary outcome of interest might be rare or only observed after years of delay (e.g., a five-year survival rate). In such cases, it is common to use the effect of an intervention on surrogate outcomes as a proxy for its effect on long-term outcomes. Recently, Yang et al. (2020) proposed how to use surrogate outcomes to impute the missing long-term outcomes and use the imputed long-term outcomes to optimize a targeting policy; they demonstrate their approach in the context of proactive churn management. The use of short-term proxies to estimate long-term treatment effects is discussed in more detail by Athey et al. (2016).

In practice, it is common for firms to make targeted intervention decisions without any causal modeling at all. Instead, in large-scale predictive applications such as ad targeting and churn incentive targeting, practitioners typically target based on models predicting outcomes (e.g., whether someone will buy after being shown an ad) rather than models predicting treatment effects (Ascarza et al. 2018). For example, in targeted online advertising, although ad campaigns are sometimes evaluated in terms of their causal effect (e.g., through A/B testing), causal effect models are rarely used for actually targeting the ads. Instead, potential customers are usually targeted based on predictions of how likely they are to convert, which are biased estimates of individual causal effects because they do not account for the counterfactual (e.g., a customer may buy even without the ad).

Researchers and savvy practitioners see this as caused by naivety, lack of modeling sophistication, or pragmatics (Provost and Fawcett 2013, Ascarza et al. 2018). However, outcome prediction models are used even when practitioners have looked carefully at the estimation of treatment effects (see Stitelman et al. (2011) and Perlich et al. (2014) for a clear example). We suggest that this use of outcome models rather than causal effect models can be seen as proxy modeling—and we should not simply criticize the practitioners for lack of modeling sophistication but rather ask what George Box might have asked: are the proxy models more useful?

Although practitioners should indeed proceed with caution when targeting with outcome prediction models (Radcliffe and Surry 2011, Ascarza 2018), prior research studies do show evidence that targeting based on outcome predictions can be effective. For example, in an advertising campaign of a major food chain, Stitelman et al. (2011) found that displaying ads based on likelihood of conversion increased the volume of prospective new customers by 40% more than displaying ads to the general population. Huang et al. (2015) deployed a churn prediction model in a telecommunications firm that increased the recharge rate of potential churners by more than 50% according to an A/B test. MacKenzie et al. (2013) estimated that 35% of what consumers purchase on Amazon and 75% of what they watch on Netflix come from product recommendations based on noncausal predictive models (as we understand it). None of these targeting approaches modeled more than one potential outcome, so it would be wrong[13] to interpret the model

estimates as causal effects; however, their estimates were able to identify individuals who would be positively affected by interventions.

Consistent with the proxy-modeling view, using outcome predictions can result in more effective causal decision making than using causal effect predictions. Radcliffe and Surry (2011) provide a list of several conditions under which causal effect predictions may not produce better targeting results than outcome predictions, such as when treatment effects are particularly small, and outcomes are informative of effects. Fernández-Loría and Provost (2019a) explain why, presenting a theoretical analysis comparing the two approaches and showing that larger sampling variance in treatment effect estimation may lead to more decision-making errors than the (biased) outcome prediction approach. Moreover, they show that the often stronger signal-to-noise ratio in outcome prediction may help with intervention decisions when outcomes and treatment effects are correlated. Thus, it may be possible to make good targeting decisions in some settings without actually having any data on how people behave when targeted.

We illustrate this last point using data made available by Criteo (an advertising platform) based on randomly targeting advertising to a large sample of users (Diemert et al. 2018).[14] We use "conversions" as the outcome of interest. The data include 13,979,592 instances, each representing a user with 11 features, a treatment indicator for the ad, and the label (i.e., whether the user converted or not). The treatment rate is 85%.

We consider two different targeting models: a causal tree (Athey and Imbens 2016) that predicts the CATE of the ad on conversions and a noncausal decision tree that predicts the probability of conversion when no ad is shown. Both models were trained and tuned with cross-validation using 80% of the sample (the training set), but we only used the data for the untreated instances to train and tune the noncausal tree. Therefore, the noncausal tree was trained with no data on how people behave when treated and substantially fewer data overall (recall that the treatment rate is 85%).

We then evaluated the two models on the remaining 20% of the sample (the test set) by using their predictions to score individuals and plotting the expected conversion rate as a function of the percentage of the individuals with the largest scores who are targeted, as shown in Figure 3. More specifically, because the data were collected through a randomized experiment, we can estimate the conversion rate produced by a targeting model as

$$E[D] \times E[Y \mid T = 1, D = 1] + (1 - E[D])$$
$$\times E[Y \mid T = 0, D = 0], \tag{5}$$

where $D$ is a binary variable that indicates whether the evaluated model intends to target the individual, $T$ is a binary variable that indicates whether the individual was treated in the data, and $Y$ is a binary variable that indicates whether the individual converted. Therefore, Figure 3 estimates the conversion rates that would result from using the evaluated models to decide which individuals to target, as the models target more individuals.

Surprisingly, the (noncausal) decision tree is better than the causal tree at identifying the candidates for whom the ad is most effective. The explanation behind this counterintuitive result is that features predictive of conversions are also predictive of treatment effects. However, the signal-to-noise ratio is higher for conversions than for treatment effects. Therefore, it is easier for machine learning algorithms to discriminate individuals according to outcomes than according to treatment effects. Critically, the training data set used to estimate the causal tree is more than six times larger than the training data set used to estimate the decision tree. Thus, although (in theory) the causal tree should outperform the decision tree with more data, acquiring enough training data to reach that point may be overly expensive in this case (because the data are collected by randomly targeting and withholding ads).

That said, outcome prediction models also have been shown to perform poorly in some cases, such as in customer retention applications where modeling negative effects is important (Radcliffe and Surry 2011, Ascarza 2018).

Generally, outcome prediction can be a reasonable alternative to causal effect estimation when (Fernández-Loría and Provost 2019a):

1. Outcomes and effects are correlated. For instance, Radcliffe and Surry (2011) find that marketing interventions often have the most positive effect on high-spending customers, and as a result, modeling causal effects does not benefit decision making as much in retail environments.

2. Outcomes are easier to estimate than causal effects, which may occur when data for one or more treatment conditions are limited (e.g., because of experimentation costs) or the signal-to-noise ratio is much larger for outcomes than for effects (as in our example).

3. Predictions are used to rank individuals. For example, if the goal is to intervene on the top-percent individuals with the largest effects, we only need to score and rank the individuals; whether the scores represent effect estimates is irrelevant.

The conditions in this list are interrelated. For example, when outcomes are much easier to estimate than causal effects, a moderate correlation between outcomes and effects could be enough for an outcome prediction model to outperform a causal effect model.

**Figure 3.** (Color online) Increase in Conversions Produced by a Noncausal Model and a Causal Model
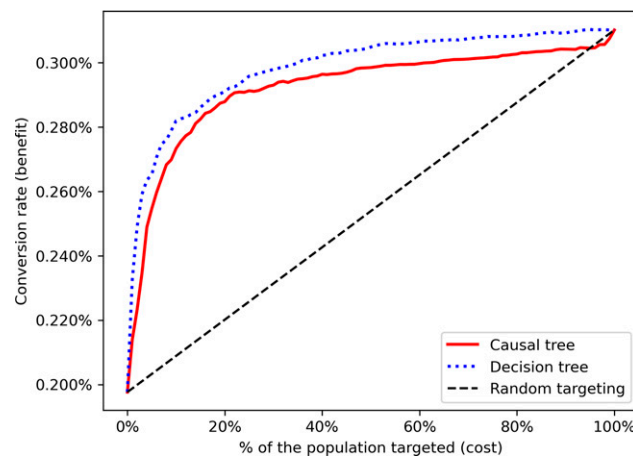


*Note.* In this specific example, the noncausal model (the decision tree) can produce at least as much benefit as the causal model for any given cost.

Similarly, a high correlation between outcomes and effects implies that ranking individuals according to outcomes or causal effects would yield similar results, so the ability to estimate outcomes better than causal effects is not as relevant. These observations apply to the specific case where the models are used to rank and then intervene on individuals (the third condition), so whether outcome prediction models could be useful for other causal tasks is an open question.

Outcome prediction is only one possible proxy task for CDM, but it is by far the most commonly used (consciously or not). For additional practical advice on when to deploy a causal effect prediction model instead of an outcome prediction model, see Radcliffe and Surry (2011).

## 4. Conclusions and Future Research

This paper develops the perspective that causal decision making and causal effect estimation are fundamentally different tasks. They are different tasks because their estimands of interest are different. Causal effect estimation assesses the impact of an intervention on an outcome of interest, and often we want or need accurate (and unbiased) estimations. In contrast, unbiased and accurate estimation of causal effects is not necessary for causal decision making, where the goal is to decide what treatment is best for each individual—a causal classification problem.

We present implications of making this distinction on (1) choosing the algorithm used to learn treatment assignment policies, that is, causal predictive models, (2) choosing the training data used to train the causal models, and (3) choosing the target variable for training the causal models. Importantly, we highlight that traditionally "good" estimates of causal effects are not necessary to make good causal decisions.

In a recent special issue of *Observational Studies* (Mitra 2021), thought leaders from across the globe wrote comments discussing Leo Breiman's influential paper "Statistical Modeling: The Two Cultures" (Breiman 2001). In his visionary paper, Breiman argued for an algorithmic modeling culture that should validate statistical models based on their prediction accuracy. When that paper was published, some individuals deemed this prediction-based focus incompatible with causal modeling (Cox 2001), but new commentaries suggest that this perspective is not at odds when "guided by a formal causal model for identification and bias reduction" (Pearl 2021, p. 188) and the algorithms incorporate "economic causal restrictions and non-prediction objectives" (Imbens and Athey 2021, p. 130).

We take a step further and argue that defining an appropriate prediction accuracy (evaluation) measure is perhaps even more important to integrate causal modeling with an algorithmic modeling culture. Both CEE and CDM can be, after all, predictive tasks. CEE can be used to predict what would happen under a counterfactual scenario, whereas CDM consists of predicting the best (counterfactual) course of action. The models can have a good CDM performance even when traditional causal assumptions are violated or not incorporated by the models.

Of course, this does not imply that firms should stop investing in randomized experiments or that causal effect estimation is not relevant for decision making. The argument here is that causal effect estimation is not necessary for *doing* effective treatment assignment. However, causal effect estimation is certainly important for the *evaluation* of causal decisions, particularly if there are concerns about how and why the decisions were made. Therefore, this paper should not be interpreted as a discourse against randomized experiments or methods for causal effect estimation.

We hope this paper will encourage practitioners to run randomized experiments evaluating all sorts of treatment assignment policies, including policies based on noncausal statistical models. Furthermore, randomized experiments can be used to identify the underlying theoretical mechanisms that result in some treatments being more effective for certain subpopulations (Tafti and Shmueli 2020), which may lead to the development of even more effective treatments.

There are multiple major opportunities for future research in causal decision making. First, most studies that distinguish between causal decision making and causal effect estimation do so in the context of decisions that are categorical and independent. This scope limitation has allowed researchers to draw parallels between causal decision making and classification. However, few studies have discussed continuous decisions (e.g., Dubé and Misra 2017) or settings where multiple treatments are assigned to the same individuals over time (Chakraborty and Murphy 2014). The results discussed in this paper imply that what matters in causal decision making is the *ranking* of the alternatives available to the decision maker rather than accurately estimating the outcomes or effects associated with each alternative. Therefore, prediction errors that preserve the ranking should not affect continuous or time-dependent decisions either. A unified and general framework explaining how prediction errors may affect causal decision making in broader settings is thus a natural step for future research.

Another promising direction for future work is the mathematical formalization of the conditions under which optimal treatment assignments can be learned from data. In causal effect estimation, several studies have proposed mathematical assumptions to characterize when causal effects can be consistently estimated from data. Well-known examples include SUTVA (Cox 1958), ignorability (Rosenbaum and Rubin 1983), back-door criterion (Pearl 2009), front-door criterion (Pearl 2009), transportability (Pearl and Bareinboim 2011), and other assumptions for the identification of causal effects using instrumental variables (Angrist et al. 1996). The perspective presented in this paper suggests that such assumptions are not necessary to learn optimal treatment assignments from data. As an example, assuming the unconfoundedness assumption is violated and the selection mechanism producing the confounding is a function of the causal effect—so that the larger the causal effect the stronger the selection—then (intuitively) the ranking of the preferred treatment alternatives should be preserved in the confounded setting, allowing the estimation of optimal treatment assignment policies from data. Mathematically formalizing such settings and linking them to real-world applications is a promising direction for future research.

The results discussed here also suggest that machine learning procedures that combine multiple sources of data (e.g., observational and experimental data) could conceivably lead to better intervention decisions by optimizing the tradeoff between confounding bias and supervised learning errors. Hence, another promising area for future research in causal decision making is the development of methods that can combine multiple sources of data. Several pioneering efforts to combine observational and experimental data for causal inference have been made (Peysakhovich and Lada 2016, Kallus et al. 2018, Athey et al. 2020, Fernández-Loría and Provost 2020, Rosenman et al. 2020), but most have the intention of estimating causal effects (rather than treatment assignment policies). Another interesting direction for the development of new methods is to try "clever" experimentation. For instance, decision makers could focus on running randomized experiments in parts of the feature space where the confounding is particularly hurtful for decision making, resulting in higher returns on their experimentation budget. Recent efforts seeking to maximize the decision-making benefits of randomized controlled trials while minimizing experimentation costs (Feit and Berman 2019, Miller and Berman 2020, Simester et al. 2020) suggest that this is a particularly promising research direction.

## Endnotes

[1] We hope that one contribution of this article is the collection of recent articles on the topic we point to.

[2] Also compare this with distinctions between explanatory modeling and predictive modeling (Shmueli 2010).

[3] The threshold for the decision could be a value other than zero, for example, when treatment costs and benefits are taken into account; see the following discussion.

[4] Few firms have the ability to deploy full-blown online machine learning systems that can manage the exploration/exploitation tradeoff dynamically. It is much more common to deploy the learned models/prediction systems than the machine learning systems themselves.

[5] This allows what follows to apply when more than only one possible treatment exists.

[6] In this case, the task was to build Spotify playlists that will lead to longer listening.

[7] We explicitly call this "causal" bias to distinguish it from all the other types of bias involved in statistical estimation. In particular, this bias is different from the statistical bias usually examined in the context of machine learning's bias/variance tradeoff. A separate causal bias/variance tradeoff exists (Fernández-Loría and Provost 2019b).

[8] Understanding how confounding affects intervention decision making is useful even when experiments are feasible. For example, as we describe next, causal bias may either help or hurt decision making. Careful experimentation may help to determine whether it would be profitable to invest in correcting causal bias.

[9] If $\tau$ had been closer to $\delta$, the UM would make the wrong decision more often.

[10] More generally, it happens when the bias pushes the estimation further in the direction of the correct decision.

[11] The binary treatment distinction is purely for ease of illustration. With multiple treatment options, we have a multiclass classification problem. As mentioned earlier, causal decision making can be reduced to a classification task where individuals should be classified into the treatment level (class) that produces the most beneficial outcome (Zadrozny 2003, Zhao et al. 2012).

[12] The threshold could be based on a cost/benefit matrix (Provost and Fawcett 2013), a budget, a workforce constraint (Provost and Fawcett 2001), or even instance specific characteristics (Saar-Tsechansky and Provost 2007).

[13] Except under extreme assumptions.

[14] See https://ailab.criteo.com/criteo-uplift-prediction-dataset/ for details and access to the data. We use the version of the data set without leakage.

## References

Angrist J, Imbens G, Rubin D (1996) Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91(434): 444–455.

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

Ascarza E, Neslin S, Netzer O, Anderson Z, Fader P, Gupta S, Hardie B, et al (2018) In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs Solutions* 5(1-2):65–81.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc. National Acad. Sci. USA* 113(27):7353–7360.

Athey S, Imbens GW (2017) The state of applied econometrics: Causality and policy evaluation. *J. Econom. Perspectives* 31(2):3–32.

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. *Annu. Rev. Econom.* 11:685–725.

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.

Athey S, Chetty R, Imbens G, Kang H (2016) Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. Preprint, submitted March 30 (v1), last revised February 29, 2020 (v3), https://arxiv.org/abs/1603.09326.

Beygelzimer A, Langford J (2009) The offset tree for learning with partial labels. *Proc.15th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 129–138.

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

Breiman L (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16(3):199–231.

Chakraborty B, Murphy SA (2014) Dynamic treatment regimes. *Annu. Rev. Statist. Appl.* 1:447–464.

Cox DR (1958) *Planning of Experiments* (Wiley, New York).

Cox DR (2001). Statistical modeling: The two cultures. *Statist. Sci.* 16(3):216–218.

Dalessandro B, Hook R, Perlich C, Provost F (2015) Evaluating and optimizing online advertising: Forget the click, but there are good proxies. *Big Data* 3(2):90–102.

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

Dorie V, Hill J, Shalit U, Scott M, Cervone D (2019) Automated vs. do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* 34(1):43–68.

Dubé JP, Misra S (2017) Personalized pricing and customer welfare. Preprint, submitted June 26, https://dx.doi.org/10.2139/ssrn.2992257.

Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize". *Management Sci.*, ePub ahead of print March 12, https://doi.org/10.1287/mnsc.2020.3922.

Elmachtoub A, Liang JCN, McNellis R (2020) Decision trees for decision-making under the predict-then-optimize framework. *Internat. Conf. on Machine Learning* (IMLS), 2858–2867.

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

Fernández-Loría C, Provost F (2019a) Causal classification: Treatment effect vs. outcome prediction. Preprint, submitted June 26, https://dx.doi.org/10.2139/ssrn.3408524.

Fernández-Loría C, Provost F (2019b) Observational vs experimental data when making automated decisions using machine learning. Preprint, submitted September 5, https://dx.doi.org/10.2139/ssrn.3444678.

Fernández-Loría C, Provost F (2020) Combining observational and experimental data to improve large-scale decision-making. *Proc. Internat. Conf. on Inform. Systems* (Association for Information Systems, Atlanta), 1583.

Fernández-Loría C, Provost F, Anderton J, Carterette B, Chandar P (2020) A comparison of methods for treatment assignment with an application to playlist generation. Preprint, submitted April 24, 2020 (v1), last revised April 23, 2021 (v4), https://arxiv.org/abs/.

Friedman JH (1997) On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining Knowledge Discovery* 1(1):55–77.

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. Graphics Statist.* 20(1):217–240.

Huang Y, Zhu F, Yuan M, Deng K, Li Y, Ni B, Dai W, et al (2015) Telco churn prediction with big data. *Proc. ACM SIGMOD Internat. Conf. on Management of Data* (ACM, New York), 607–618.

Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist.* 7(1): 443–470.

Imbens G, Athey S (2021) Breiman's two cultures: A perspective from econometrics. *Observational Stud.* 7(1):127–133.

Kallus N, Puli AM, Shalit U (2018) Removing hidden confounding by experimental grounding. *Proc. 32nd Internat. Conf. on Neural Inform. Processing Systems, NeurIPS* (Neural Information Processing Systems Foundation, California), 10911–10920.

Lemmens A, Gupta S (2020) Managing churn to maximize profits. *Marketing Sci.* 39(5):956–973.

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proc. 19th Internat. Conf. on World Wide Web* (ACM, New York), 661–670.

MacKenzie I, Meyer C, Noble S (2013) *How Retailers Can Keep Up with Consumers* (McKinsey & Company).

Manski CF (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–1246.

McFowland III E, Somanchi S, Neill DB (2018) Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. Preprint, submitted March 24 (v1), last revised June 7, 2018 (v2), https://arxiv.org/abs/1803.09159.

McFowland III E, Gangarapu S, Bapna R, Sun T (2022) *A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects. MIS Quart.* Forthcoming.

Miller A, Berman R (2020) Test, target, & roll: Optimal explore-first contextual targeting in finite populations. *Poster Session, Conference on Digital Experimentation* (CODE@MIT, Boston).

Miller A, Hosanagar K (2020) Personalized discount targeting with causal machine learning. *Proc. Internat. Conf. on Inform. Systems* (Association for Information Systems, Atlanta), 1682.

Mitra N (2021) Introduction. *Observational Stud.* 7(1):1–2.

Olaya D, Coussement K, Verbeke W (2020) A survey and benchmarking study of multitreatment uplift modeling. *Data Mining Knowledge Discovery* 34(2):273–308.

Pearl J (2009) *Causality: Models, Reasoning and Inference* (Cambridge University Press).

Pearl J (2021) Causally colored reflections on leo breiman's "statistical modeling: The two cultures" (2001). *Observational Stud.* 7(1):187–190.

Pearl J, Bareinboim E (2011) Transportability of causal and statistical relations: A formal approach. *Proc. AAAI Conf. on Artificial Intelligence AAAI*, vol. 25 (Advancement of Artificial Intelligence, California).

Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F (2014) Machine learning for targeted display advertising: Transfer learning in action. *Machine Learn.* 95(1):103–127.

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.

Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. *Statist. Medicine* 8(4):431–440.

Provost F, Fawcett T (2001) Robust classification for imprecise environments. *Machine Learn.* 42(3):203–231.

Provost F, Fawcett T (2013) *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media).

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 E, Basse G, Owen A, Baiocchi M (2020) Combining observational and experimental datasets using shrinkage estimators. Preprint, submitted February 16 (v1), last revised May 18, 2020 (v2), https://arxiv.org/abs/2002.06708.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psych.* 66(5):688.

Saar-Tsechansky M, Provost F (2007) Decision-centric active learning of binary-outcome models. *Inform. Systems Res.* 18(1):4–22.

Schuler A, Baiocchi M, Tibshirani R, Shah N (2018) A comparison of methods for model selection when estimating individual treatment effects. Preprint, submitted April 14 (v1), last revised June 13, 2018 (v2), https://arxiv.org/abs/1804.05146.

Shmueli G (2010) To explain or to predict? *Statist. Sci.* 25(3):289–310.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Sci.* 66(8):3412–3424.

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations Trends Machine Learn.* 12(1-2):1–286.

Stitelman O, Dalessandro B, Perlich C, Provost F (2011) Estimating the effect of online display advertising on browser conversion. *Proc. Fifth International Workshop on Data Mining and Audience Intelligence for Advertising, ADKDD* (ACM, New York), 8–16.

Tafti A, Shmueli G (2020) Beyond overall treatment effects: Leveraging covariates in randomized experiments guided by causal structure. *Inform. Systems Res.* 31(4):1183–1199.

VanderWeele TJ (2013) Surrogate measures and consistent surrogates. *Biometrics* 69(3):561–565.

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* (Nelson Education).

Yahav I, Shmueli G, Mani D (2015) A tree-based approach for addressing self-selection in impact studies with big data. *Management Inform. Systems Quart.* 40(4):819–848.

Yang J, Eckles D, Dhillon P, Aral S (2020) Targeting for long-term outcomes. Preprint, submitted October 29, https://arxiv.org/abs/2010.15835.

Zadrozny B (2003) Policy mining: Learning decision policies from fixed sets of data. PhD thesis, University of California, San Diego.

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.