Semiparametric Estimation of Long-Term Treatment Effects*

Jiafeng Chen
Harvard Business School
jiafengchen@g.harvard.edu

David M. Ritzwoller
Stanford Graduate School of Business
ritzwoll@stanford.edu

ABSTRACT. Long-term outcomes of experimental evaluations are necessarily observed after long delays. We develop semiparametric methods for combining the short-term outcomes of experiments with observational measurements of short-term and long-term outcomes, in order to estimate long-term treatment effects. We characterize semiparametric efficiency bounds for various instances of this problem. These calculations facilitate the construction of several estimators. We analyze the finite-sample performance of these estimators with a simulation calibrated to data from an evaluation of the long-term effects of a poverty alleviation program.

Keywords: Long-Term Treatment Effects, Semiparametric Efficiency, Experimentation
JEL Codes: C01, C13, C14, O10

1. Introduction

Empirical researchers often aim to estimate the long-term effects of policies or interventions. Randomized experimentation provides a simple and interpretable approach to this problem (Athey and Imbens, 2017; Duflo et al., 2007; Fisher, 1925). However, long-term outcomes of experimental evaluations are necessarily observed after long and potentially costly delays. Consequently, there is relatively limited experimental evidence on the long-term effects of economic and social policies.[1]

In this paper, we develop methods for estimating long-term treatment effects by combining short-term experimental and long-term observational data sets. We consider two closely related models, proposed by Athey et al. (2020a) and Athey et al. (2020b). In both cases, a researcher is interested in

Date: August 17, 2023

*We thank Isaiah Andrews, Xiaohong Chen, Raj Chetty, Pascaline Dupas, Bryan Graham, Han Hong, Guido Imbens, Melanie Morten, Evan Munro, Hyunseung Kang, Brad Ross, Joseph Romano, Suproteem Sarkar, Rahul Singh, Elie Tamer, Stefan Wager, and Chris Walker for helpful comments and conversations. Ritzwoller gratefully acknowledges support from the National Science Foundation under the Graduate Research Fellowship.

[1] Bouguen et al. (2019) provide a systematic review of randomized control trials in development economics and report that only a small proportion evaluate long-term effects. Similarly, in a review on experimental evaluations of the effects of early childhood educational interventions, Tanner et al. (2015) identify only one randomized evaluation that reports long-term employment and labor market effects (Gertler et al., 2014).

the average effect of a binary treatment on a scalar long-term outcome. The researcher observes two samples of data, an *experimental* sample and an *observational* sample. The experimental sample measures short-term outcomes of a randomized evaluation of the treatment. the treatment The observational sample measures both short-term and long-term outcomes, but may be subject to unmeasured confounding. The two settings that we consider are distinguished by whether treatment is observed in the observational sample. Similar identifying assumptions are required in each case. We review these assumptions in Section 2.

Methods for combining experimental and observational data to estimate long-term treatment effects are widely applicable in applied microeconomics and online platform experimentation. The estimators proposed in Athey et al. (2020b) have been used to estimate the long-term effects of free tuition on college completion (Dynarski et al., 2021), of an agricultural monitoring technology on farmer revenue in Paraguay (Dal Bó et al., 2021), and of changes to Twitter's platform on user engagement (Twitter Engineering, 2021). As a result, there is considerable interest in advancing a statistical and methodological foundation for this problem (Gupta et al., 2019).

To that end, this paper offers two contributions. First, we develop semiparametric theory for estimation of long-term average treatment effects in Section 3. In particular, we derive the semiparametric efficiency bound, and the corresponding efficient influence function, for estimating long-term average treatment effects in each of the models that we consider. We then demonstrate that the efficient influence function is the unique influence function in each model, indicating that all regular and asymptotically linear estimators have the same asymptotic variance and achieve the semiparametric efficiency bound (Chen and Santos, 2018; Newey, 1994). In both cases, we find that the efficient influence functions possess a "double-robust" structure commonly found in causal estimation problems (Kang and Schafer, 2007; Scharfstein et al., 1999).

These results are novel. In particular, our calculations correct statements concerning efficient influence functions and semiparametric efficiency bounds given in a working paper draft of Athey et al. (2020b).[2] Analogous results for the model considered in Athey et al. (2020a) have not appeared before in the literature.

Second, in Section 4, we establish the consistency and asymptotic normality of a suite of estimators. These estimators differ according to whether they are based on moment conditions associated with the efficient influence functions derived in Section 3. Moment conditions defined by efficient influence functions are often referred to as Neyman orthogonal moment conditions, due

[2] The results given in Athey et al. (2020b) were obtained through a non-rigorous calculation related to standard heuristics involving a discretization of the sample space (see e.g., Kennedy (2022) and Ichimura and Newey (2022) for discussion). Our arguments are rigorous, following the method developed in Section 3.4 of Bickel et al. (1993).

to their insensitivity to local perturbations of nuisance parameters (see e.g., Chernozhukov et al., 2018; Foster and Syrgkanis, 2019). The estimators that we formulate can be viewed as instances of one-step estimators (Bickel, 1982; Le Cam, 1956; Newey, 1994). Roughly, the conditional expectations that compose an efficient influence function are estimated by augmenting standard machine learning algorithms with cross-fitting. Estimates of long-term average treatment effects are then formed by treating estimated efficient influence functions as identifying moment functions. Estimators obtained from Neyman orthogonal moment conditions admit a very general analysis, under high-level sufficient conditions, due to Chernozhukov et al. (2018). We adapt these arguments to our setting.

We compare these estimators to a variety of alternative estimators based on non-orthogonal moment conditions. In particular, we consider a set of estimators that are analogous to standard inverse propensity score weighting and outcome regression estimators for average treatment effects under unconfoundedness (see e.g., Imbens 2004 for a review). These estimators include, but are not limited to, many of the estimators proposed in Athey et al. (2020a) and Athey et al. (2020b), and applied by e.g., Dynarski et al. (2021). Here, to obtain theoretical guarantees, we restrict attention to estimators that plug-in nuisance parameter estimates derived from the method of sieves (Chen and Liao, 2015; Chen et al., 2014). Relative to those for estimators based on orthogonal moments, sufficient conditions for estimators based on non-orthogonal moments appear more stringent.

Section 5 assesses the finite sample performance of these semiparametric estimators with a simulation calibrated to data from a randomized evaluation of the long-term effects of a poverty alleviation program originally analyzed in Banerjee et al. (2015). We find that estimators based on orthogonal moments are substantially more accurate than estimators based on non-orthogonal moments. Section 6 concludes.

Proofs for all results stated in the main text are provided in Appendix A. Appendices B to D give additional results or details and will be introduced at appropriate points throughout the paper. Code implementing the estimators developed in this paper is available on GitHub at the link https://github.com/DavidRitzwoller/longterm.

1.1 Related literature. Settings closely related to our own include Rosenman et al. (2018), Rosenman et al. (2020), and Kallus and Mao (2020). In Rosenman et al. (2018), treatment assignment is unconfounded in both samples. In Rosenman et al. (2020), the outcome of interest is observed in both samples. In Kallus and Mao (2020), treatment is unconfounded in both samples considered jointly, but not necessarily in either sample considered individually. Yang et al. (2020), Gui (2020), and Gechter and Meager (2021) give complementary analyses of related models. Hou et al. (2021) consider a model in which measurements of long-term outcomes and treatments are missing completely at random. Imbens et al. (2022) propose several alternative identification strategies for models similar to the model that we consider. In contemporaneous work, Singh (2021) and Singh (2022) propose estimators and confidence intervals similar to those developed in Section 4 for the case in which treatment is not observed in the observational data set.

This paper contributes to the literature on missing data models (Hotz et al., 2005; Little and Rubin, 2019; Ridder and Moffitt, 2007), and more specifically to the literature on semiparametric efficiency in missing data models (Bia et al., 2020; Chen et al., 2008; Graham, 2011; Muris, 2020). The models that we consider are related to the literatures on statistical surrogacy (Begg and Leung, 2000; Prentice, 1989) and mediation analysis (Imai et al., 2010; van der Laan and Petersen, 2004).

1.2 Notation. Let λ be a σ-finite measure on the measurable space $(\Omega, \mathcal{F})$, and let $\mathcal{M}_\lambda$ be the set of all probability measures on $(\Omega, \mathcal{F})$ that are absolutely continuous with respect to λ. For an arbitrary random variable D defined on $(\Omega, \mathcal{F}, Q)$ with $Q \in \mathcal{M}_\lambda$, we let $\mathbb{E}_Q[D]$ denote its expected value. The quantity $p(d \mid E)$ will denote the density of the random variable D at d conditional on the event $E \in \Omega$ and $\ell(d \mid E)$ will denote the corresponding log-likelihood. This notation leaves the law of the random variable D implicit, but will cause no ambiguity. We let $\|\cdot\|_{Q,q}$ denote the $L^q(Q)$ norm and $[n]$ denote the set $\{1, \ldots, n\}$.

2. PROBLEM FORMULATION

Consider a researcher who conducts a randomized experiment aimed at assessing the effects of a policy or intervention. For each individual in the experiment, they measure a q-vector X_i of pre-treatment covariates, a binary variable W_i denoting assignment to treatment, and a d-vector S_i of short-term post-treatment outcomes. The researcher is interested in the effect of the treatment on a scalar, long-term, post-treatment outcome Y_i that is not measured in their experimental data. They are able to obtain an auxiliary, observational data set containing measurements, for a separate population of individuals, of the long-term outcome of interest, in addition to records of the same pre-treatment covariates and short-term outcomes that were measured in the experimental data set. This observational data set may or may not record whether each individual was exposed to the treatment of interest. In this paper, we develop methods for estimating the effect of a treatment on long-term outcomes that combine experimental and observational data sets with this structure.

To fix ideas, consider Dynarski et al. (2021), who randomize grants of free college tuition to a population of high achieving, low income high school students. They estimate the average effects of these free tuition grants on college application and enrollment rates. The long-term effects of free tuition grants on college completion rates may be of more direct interest to policy-makers considering the expansion of college aid programs. However, it will take several years before college completion is observed for the cohort of students in the experimental sample. Given an observational data set that records college enrollment and completion for a comparable population of high school students, the methods developed in this paper may facilitate a more timely quantification of the effect of free tuition grants on college completion.

We begin this section by defining the data structure and estimands that we consider. We then review, and restate in a common notation, two sets of closely related sets of identifying assumptions proposed by Athey et al. (2020a) and Athey et al. (2020b). We refer to these settings as the *Latent Unconfounded Treatment* and *Statistical Surrogacy* Models, respectively.

2.1 Data. Consider the collection of random variables

$$\{A_i\}_{i=1}^n = \{(Y_i(0), Y_i(1), S_i(0), S_i(1), W_i, G_i, X_i)\}_{i=1}^n$$

consisting of the potential outcomes and characteristics of a sample of individuals drawn independently and identically from a distribution $P_\star$. Here, G_i is a binary indicator denoting whether the observation was acquired in the observational sample ($G_i = 1$) or the experimental sample ($G_i = 0$). The variables $Y_i(\cdot)$ are long-term potential outcomes, $S_i(\cdot)$ are short-term potential outcomes, W_i is a binary treatment indicator, and X_i are covariates.

The data observable to the researcher are denoted by $\{B_i\}_{i=1}^n$ and are i.i.d. according to a distribution denoted by P. The *observed* outcomes S_i and Y_i are given by $S_i = W_i S_i(1) + (1 - W_i)S_i(0)$ and $Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$, respectively. The short-term outcomes S_i are observed in both the observational and experimental data sets. The long-term outcome Y_i is observed only in the observational data set. Treatment W_i may or may not be observed in the observational sample. Thus, the observable data B_i are given by $(G_i Y_i, S_i, W_i, G_i, X_i)$ if treatment is measured in the observational data set and by $(G_i Y_i, S_i, (1 - G_i)W_i, G_i, X_i)$ if treatment is not measured in the observational data set.

2.2 Estimands. In the main text, we consider estimation of the long-term average treatment effect in the observational population, given by

$$\tau_1 = \mathbb{E}_{P_\star}\left[Y_i(1) - Y_i(0) \mid G_i = 1\right] . \tag{2.1}$$

Athey et al. (2020a) note that there is often reason to believe that features of the observational population ($G_i = 1$) are more "externally valid," in that they of greater interest to policymakers.

In Appendix B, we give results analogous to those presented in the main text for the long-term average treatment effect in the experimental population

$$\tau_0 = \mathbb{E}_{P_\star}\left[Y_i(1) - Y_i(0) \mid G_i = 0\right] \ , \tag{2.2}$$

which may be of interest in some contexts.[3] Athey et al. (2020a) and Athey et al. (2020b) consider estimation of τ_1 and τ_0, respectively.

2.3 Identifying Assumptions. We consider two sets of assumptions, proposed in Athey et al. (2020a) and Athey et al. (2020b). In both models, the long-term average treatment effect τ_1 is identified. The models differ according to whether they are applicable to contexts in which treatment is or is not measured in the observational data set. Both models assume that treatment assignment is unconfounded in the experimental data set and that the probabilities of being assigned treatment or of being measured in the observational data set satisfy a strict overlap condition.

Assumption 2.1 (Experimental Unconfounded Treatment)**.** *In the experimental data set, treatment is independent of short-term and long-term potential outcomes conditional on pre-treatment covariates, in the sense that*

$$W_i \perp\!\!\!\perp (Y_i(0), S_i(0), Y_i(1), S_i(1)) \mid X_i, G_i = 0 \ .$$

Assumption 2.2 (Strict Overlap)**.** *The probability of being assigned to treatment or of being measured in the observational data set is strictly bounded away from zero and one, in the sense that, for each w and g in $\{0, 1\}$, the conditional probabilities*

$$P(W = w \mid S, X, G = g) \quad \textit{and} \quad P(G = g \mid S, X, W = w)$$

are bounded between ε and $1 - \varepsilon$, λ-almost surely, for some fixed constant $0 < \varepsilon < 1/2$.

Unconfounded treatment and strict overlap are often satisfied in the experimental sample by design. Assessing and accounting for violations of overlap in observational data are important empirical and methodological issues (Crump et al., 2009). In particular, strong overlap conditions can place stringent restrictions on the data generating process when there are many covariates or short-term outcomes (D'Amour et al., 2021). We view systematic consideration of these issues in our context as an important area for further research.

In addition, as our aim is to use the experimental sample to estimate a feature of the observational population, we require an assumption limiting the differences between the two populations.

[3] An alternative estimand is the unconditional long-term average treatment effect $\tau = \mathbb{E}_{P_\star}\left[Y_i(1) - Y_i(0)\right]$. The practical interpretation of this parameter is somewhat nebulous, as it is unclear why it would be desirable to weight the two samples in the definition of the parameter according to their sizes.

Assumption 2.3 (Experimental Conditional External Validity)**.** *The distribution of the potential outcomes is invariant to whether the data belong to the experimental or observational data sets, in the sense that*

$$G_i \perp\!\!\!\perp (Y_i(1), Y_i(0), S_i(1), S_i(0)) \mid X_i \ .$$

Assumption 2.3 implies that adjustments to the distribution of covariates in the experimental data set are sufficient to obtain approximations to features of the observational population, thereby ruling out unobserved systematic differences between the two populations conditional on covariates.

2.3.1 Latent Unconfounded Treatment. If treatment is measured in the observational data set, then the key identifying assumption is that treatment assignment is unconfounded with respect to the long-term outcome if conditioned on the short-term potential outcomes. We term this restriction "Latent Unconfounded Treatment" following Athey et al. (2020a).

Assumption 2.4 (Observational Latent Unconfounded Treatment)**.** *In the observational data set, treatment is independent of the long-term potential outcomes conditional on the short-term potential outcomes and pre-treatment covariates, in the sense that, for* $w \in \{0, 1\}$,

$$W_i \perp\!\!\!\perp Y_i(w) \mid S_i(w), X_i, G_i = 1 \ .$$

Informally, Assumption 2.4 states that all unobserved confounding in the observational sample is mediated through the short-term outcomes. Assumptions 2.1 to 2.4 are sufficient for identification of the long-term treatment effect τ_1. We summarize these assumptions with the following shorthand.

Definition 2.1. The collection of Assumptions 2.1 to 2.4, in a setting where treatment is measured in the observational data set, is referred to as the Latent Unconfounded Treatment Model.

Panel A of Figure 1 displays a causal Directed Acyclic Graph (DAG) that is consistent with the restrictions that the Latent Unconfounded Treatment Model place on the data generating process for the observational data set.[4] The following proposition is stated as Theorem 1 in Athey et al. (2020a). We state and prove the result for completeness.

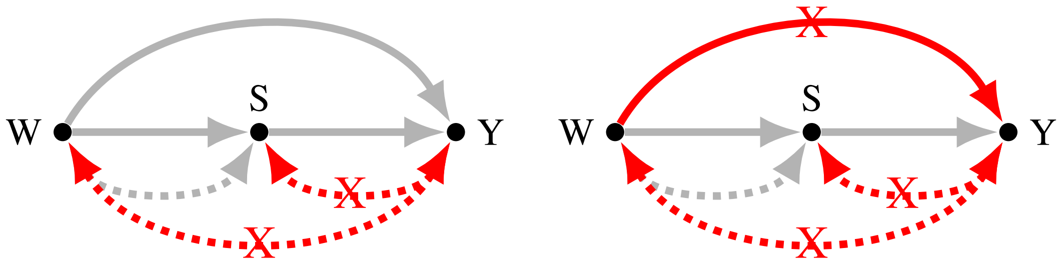
Proposition 2.1 (Athey et al. (2020a))**.** *Under the Latent Unconfounded Treatment Model, the long-term average treatment effect* τ_1 *is point identified.*

2.3.2 Statistical Surrogacy. If treatment is not measured in the observational data set, an alternative "Statistical Surrogacy" assumption, in the spirit of Prentice (1989), is required in the place of

[4] See Pearl (1995) for further discussion of the applications of manipulation of graphical models to causal inference. We do not make use of do-calculus methodology to establish identification.

FIGURE 1. Causal DAGs Consistent with Assumptions on Observational Data

Panel A: Latent Unconfounded Treatment *Panel B: Statistical Surrogacy*



Notes: Panels A and B of Figure 1 display causal DAGs that describe the restrictions on the data generating process for the observational data set ($G = 1$) implied by the Latent Unconfounded Treatment and Statistical Surrogacy Models, respectively. Light grey arrows denote the existence of an effect of the tail variable on the head variable. Dark red arrows with x's denote that an effect of the the tail variable on the head variable is ruled out. Dashed bidirectional arrows represent existence of some unobserved common causal variable U, where we have a fork $\leftarrow U \rightarrow$.

Assumption 2.4. Under this restriction, the short-term outcomes can be interpreted as "proxies" or "surrogates" for the long-term outcome.

Assumption 2.5 (Experimental Statistical Surrogacy)**.** *In the experimental data set, treatment is independent of the long-term observed outcomes conditional on the short-term observed outcomes and pre-treatment covariates, in the sense that*

$$W_i \perp\!\!\!\perp Y_i \mid S_i, X_i, G_i = 0 \ .$$

Informally, Assumption 2.5 additionally rules out a causal link from treatment to the long-term outcome that is not mediated by the short-term outcomes.

In addition to Assumption 2.3, a supplementary restriction is required to ensure that the experimental and observational data sets are suitably comparable, conditional on the observed outcomes.

Assumption 2.6 (Long-Term Outcome Comparability)**.** *The distribution of the long-term outcome is invariant to whether the data belong to the experimental or observational data sets conditional on the short-term outcome and covariates, in the sense that*

$$G_i \perp\!\!\!\perp Y_i \mid X_i, S_i \ .$$

Assumption 2.3 is not strictly stronger than Assumption 2.6, as belonging to the experimental or observational data sets is not necessarily independent of treatment assignment. Statistical Surrogacy, in addition to Assumptions 2.1 to 2.3 and 2.6, is sufficient for identification of the long-term treatment effect τ_1 in settings where treatment is not measured in the observational data set. Again, we summarize these conditions with the following shorthand.

Definition 2.2. The collection of Assumptions 2.1 to 2.3, 2.5, and 2.6, in a setting where treatment is not measured in the observational data set, is referred to as the Statistical Surrogacy Model.

Panel B of Figure 1 describes the restrictions that the Statistical Surrogacy Model place on the data generating process for the observational data set. The following proposition restates Theorem 1 of Athey et al. (2020b). Again, we state and prove the result for completeness.[5]

Proposition 2.2 (Athey et al. (2020b))**.** *Under the Statistical Surrogacy Model, τ_1 is point identified.*

3. SEMIPARAMETRIC EFFICIENCY

In this section, we derive efficient influence functions and corresponding semiparametric efficiency bounds for estimation of long-term average treatment effects τ_1 in observational populations. In Appendix B, we state comparable results for long-term average treatment effects τ_0 in experimental populations.

3.1 Nuisance Functions. The efficient influence functions that we derive are expressed in terms of a set of unknown, but identified, conditional expectations. We classify each of these objects as being either a "long-term outcome mean" or a "propensity score." Each long-term outcome mean is an expectation of the long-term outcome conditioned on other features of the data. Under the Latent Unconfounded Treatment Model, where treatment is measured in the observational data set, the long-term outcome means that appear in the efficient influence function are given by

$$\mu_w(s,x) = \mathbb{E}_P[Y \mid W = w, S = s, X = x, G = 1] \quad \text{and} \tag{3.1}$$

$$\overline{\mu}_w(x) = \mathbb{E}_P[\mu_w(S,X) \mid W = w, X = x, G = 0]\,. \tag{3.2}$$

The function $\mu_w(s,x)$ is the mean of the long term outcomes in the observational sample, conditioned on treatment w, the short-term outcomes s, and the covariates x. The function $\overline{\mu}_w(x)$ is the projection of $\mu_w(s,x)$ onto the experimental population conditional on w and x, integrated over s.

In turn, under the Statistical Surrogacy model, where treatment is not measured in the observational data set, the analogous nuisance functions appearing in the efficient influence function are given by

$$\nu(s,x) = \mathbb{E}_P[Y \mid S = s, X = x, G = 1] \quad \text{and} \tag{3.3}$$

$$\overline{\nu}_w(x) = \mathbb{E}_P[\nu(S,X) \mid W = w, X = x, G = 0]\,. \tag{3.4}$$

[5] We note that Assumption 2.3 is unnecessary for the identification of τ_0 in the Statistical Surrogacy Model.

The function $\nu(s,x)$ is similar to $\mu_w(s,x)$, but does not condition on treatment, as treatment is not observed in the observational sample under the Statistical Surrogacy Model. The function $\overline{\nu}_w(x)$ projects $\nu(s,x)$ on the observational sample, conditional on x and w, and integrates over s.

It useful to note that the proofs of Propositions 2.1 and 2.2 operate by establishing, in their respective models, that the equalities

$$\mathbb{E}_{P_*}\left[Y(1) \mid G=1\right] = \mathbb{E}_{P_*}\left[\overline{\mu}_1(X) \mid G=1\right] \quad \text{and} \quad \mathbb{E}_{P_*}\left[Y(1) \mid G=1\right] = \mathbb{E}_{P_*}\left[\overline{\nu}_1(X) \mid G=1\right]$$

hold and that the objects (3.2) and (3.4) are identified from the observable data. Section 4.2 presents a set of alternative moment conditions that analogously identify τ_1.

Each member of the second class of nuisance functions—propensity scores—expresses either the probability of treatment or the probability of inclusion in the observational sample conditioned on other features of the data. In particular, let

$$\rho_w(s,x) = P_\star(W=w \mid S(w)=s, X=x, G=1)\,, \tag{3.5}$$

$$\varrho(s,x) = P(W=1 \mid S=s, X=x, G=0)\,, \quad \text{and} \tag{3.6}$$

$$\varrho(x) = P(W=1 \mid X=x, G=0) \tag{3.7}$$

denote the probabilities of treatment conditional on various features of the data. Similarly, let

$$\gamma(s,x) = P(G=1 \mid S=s, X=x)\,, \gamma(x) = P(G=1 \mid X=x)\,, \text{ and } \pi = P(G=1) \tag{3.8}$$

denote the probabilities of inclusion in the observational sample conditional on various features of the data.[6] Each of these objects is bounded away from zero and one λ-almost surely by strict overlap (Assumption 2.2). Although the propensity score $\rho_w(s,x)$ includes the unobserved random variable $S(w)$ in its conditioning set, we may write $\rho_w(s,x)$ in terms of observables as

$$\begin{aligned}\rho_w(s,x) = {} & \frac{P(G=1 \mid S=s, W=w, X=x)}{P(G=0 \mid S=s, W=w, X=x)} \\ & \times \frac{P(G=0 \mid W=w, X=x)}{P(G=1 \mid W=w, X=x)} P(W=w \mid X=x, G=1)\end{aligned} \tag{3.9}$$

by successive applications of Bayes' rule, Assumption 2.1, and Assumption 2.3.

3.2 Influence Functions and Efficiency Bounds. We characterize efficient influence functions for τ_1 with the approach developed in Section 3.4 of Bickel et al. (1993). In particular, we characterize the tangent spaces for the classes of distributions restricted by the maintained assumptions. We then verify that the conjectured efficient influence functions are pathwise derivatives of τ_1 and

[6] Athey et al. (2020b) term $\nu(s,x)$ the *surrogate index*, $\varrho(s,x)$ the *surrogate score*, and $1-\gamma(s,x)$ the *sampling score*.

are elements of their respective tangent spaces. Recall that the Latent Unconfounded Treatment Model and Statistical Surrogacy Model differ according to the assumptions they impose on the data generating distribution $P_\star$ in addition to whether treatment is assumed to have been measured in the observational sample.

Theorem 3.1. *Let $b = (y, s, w, g, x)$ denote a possible value for the observed data.*

(1) Under the Latent Unconfounded Treatment Model, given in Definition 2.1, the efficient influence function for the parameter τ_1 is given by

$$\begin{aligned}\psi_1(b,\tau_1,\eta) = \frac{g}{\pi}\left(\frac{w(y-\mu_1(s,x))}{\rho_1(s,x)} - \frac{(1-w)(y-\mu_0(s,x))}{\rho_0(s,x)} + (\overline{\mu}_1(x)-\overline{\mu}_0(x)) - \tau_1\right) \\ + \frac{1-g}{\pi}\left(\frac{\gamma(x)}{1-\gamma(x)}\left(\frac{w(\mu_1(s,x)-\overline{\mu}_1(x))}{\varrho(x)} - \frac{(1-w)(\mu_0(s,x)-\overline{\mu}_0(x))}{1-\varrho(x)}\right)\right),\end{aligned} \tag{3.10}$$

where the parameter η collects the nuisance functions appearing in (3.10).

(2) Under the Statistical Surrogacy Model, given in Definition 2.2, the efficient influence function for the parameter τ_1 is given by

$$\begin{aligned}\xi_1(b,\tau_1,\varphi) = \frac{g}{\pi}\left(\frac{\gamma(x)}{\gamma(s,x)}\frac{1-\gamma(s,x)}{1-\gamma(x)}\frac{(\varrho(s,x)-\varrho(x))(y-\nu(s,x))}{\varrho(x)(1-\varrho(x))} + (\overline{\nu}_1(x)-\overline{\nu}_0(x)) - \tau_1\right) \\ + \frac{1-g}{\pi}\left(\frac{\gamma(x)}{1-\gamma(x)}\left(\frac{w(\nu(s,x)-\overline{\nu}_1(x))}{\varrho(x)} - \frac{(1-w)(\nu(s,x)-\overline{\nu}_0(x))}{1-\varrho(x)}\right)\right),\end{aligned} \tag{3.11}$$

where the parameter φ collects nuisance functions appearing in (3.11).[7]

In our discussion, we will frequently reference the nuisance parameters η and φ introduced in Theorem 3.1. We introduce the following notation to expedite our exposition.

Definition 3.1. Partition the nuisance function η and φ defined in Theorem 3.1 into the long-term outcome means, propensity scores, and π by

$$\eta = (\omega_\psi, \kappa_\psi, \pi) \quad \text{and} \quad \varphi = (\omega_\xi, \kappa_\xi, \pi).$$

In particular, the parameters ω_ψ and ω_ξ collect long-term outcome means

$$\omega_\psi = (\overline{\mu}_1(\cdot), \overline{\mu}_0(\cdot), \mu_1(\cdot,\cdot), \mu_0(\cdot,\cdot)) \quad \text{and} \quad \omega_\xi = (\overline{\nu}_1(\cdot), \overline{\nu}_0(\cdot), \nu(\cdot,\cdot)),$$

respectively, and the parameters κ_ψ and κ_ξ collect propensity scores

$$\kappa_\psi = (\rho_1(\cdot,\cdot), \rho_0(\cdot,\cdot), \varrho(\cdot), \gamma(\cdot)) \quad \text{and} \quad \kappa_\xi = (\varrho(\cdot,\cdot), \varrho(\cdot), \gamma(\cdot,\cdot), \gamma(\cdot)),$$

[7] We thank Rahul Singh for noting a typo in the statement of $\xi_1(\cdot)$ in a previous draft of this paper, which was missing the factor $\gamma(x)/(1-\gamma(x))$ in the second term.

respectively.

Remark 3.1. The efficient influence functions $\psi_1(b, \tau_1, \eta)$ and $\xi_1(b, \tau_1, \varphi)$ derived in Theorem 3.1 are additively separable into two terms associated with the observational and experimental data sets, respectively. The structure of the terms associated with the observational data set resembles the structure of the efficient influence function for the average treatment effect under unconfoundedness (Hahn, 1998); we discuss the relationship between these objects in Section 4.1.1. ■

Remark 3.2. The efficient influence functions $\psi_1(b, \tau_1, \eta)$ and $\xi_1(b, \tau_1, \varphi)$ possess a "double-robust" structure that is prevalent in causal inference and missing data problems (see e.g., Bang and Robins, 2005; Kang and Schafer, 2007). In particular, the mean-zero property of $\psi_1(b, \tau_1, \eta)$ is maintained even if some of the nuisance functions are misspecified. Suppose that we let arbitrary measurable functions $\tilde{\omega}$ replace the long-term outcome means ω_ψ, then $\mathbb{E}_P\left[\psi_1(B, \tau_1, (\tilde{\omega}, \kappa_\psi, \pi))\right] = 0$. Similarly, if the arbitrary measurable functions $\tilde{\kappa}$ replace the propensity scores κ_ψ, then we have that $\mathbb{E}_P\left[\psi_1(B, \tau_1, (\omega_\psi, \tilde{\kappa}, \pi))\right] = 0$. The efficient influence function $\xi_1(b, \tau_1, \varphi)$ is similarly robust to misspecification of either the long-term outcome means or the propensity scores. This result echoes an analogous double-robustness property of the efficient influence function of the average treatment effect under ignorability (Scharfstein et al., 1999), in which the efficient influence function is mean zero under misspecification of either the conditional means of the outcome variable or the propensity score. The form of the double-robustness entailed here is slightly more general, as ω and κ each collect several nuisance function.

Double robustness, in this form, has a useful implication for estimation. We demonstrate in Appendix C.1 that, under appropriate regularity conditions, estimators based on the efficient influence functions $\psi_1(b, \tau_1, \eta)$ or $\xi_1(b, \tau_1, \varphi)$ are consistent for τ_1 if *either* the outcome means (ω_ψ or ω_ξ) or the propensity scores (κ_ψ or κ_ξ) are estimated consistently. We analyze estimators of this form in further detail in Section 4. ■

The population variance of the efficient influence function is the semiparametric efficiency bound. The respective bounds are presented in Corollary 3.1.

Corollary 3.1. *Define the conditional variances*

$$\sigma_w^2(s, x) = \mathbb{E}_{P_\star}[(Y(w) - \mu_w(S, X))^2 \mid S = s, X = x] \quad \textit{and}$$
$$\sigma^2(s, x) = \mathbb{E}_{P_\star}[(Y - \nu(S, X))^2 \mid S = s, X = x]$$

as well as the expressions

$$\Gamma_{w,1}(s,x) = \frac{\gamma(x)}{1-\gamma(x)} \frac{(\mu_w(s,x) - \overline{\mu}_w(x))^2}{\varrho(x)^w (1-\varrho(x))^{1-w}} \text{ and } \Lambda_{w,1}(s,x) = \frac{\gamma(x)}{1-\gamma(x)} \frac{(\nu(s,x) - \overline{\nu}_w(x))^2}{\varrho(x)^w (1-\varrho(x))^{1-w}} .$$

(1) Under the Latent Unconfounded Treatment Model, given in Definition 2.1, the semiparametric efficiency bound for τ_1 is given by

$$V_1^{\star} = \mathbb{E}_P \left[\frac{\gamma(X)}{\pi^2} \left(\frac{\sigma_1^2(S,X)}{\rho_1(S,X)} + \frac{\sigma_0^2(S,X)}{\rho_0(S,X)} \right.\right. \\ \left.\left. + (\overline{\mu}_1(X) - \overline{\mu}_0(X) - \tau_1)^2 + \Gamma_{0,1}(S,X) + \Gamma_{1,1}(S,X) \right) \right] . \tag{3.12}$$

(2) Under the Statistical Surrogacy Model, given in Definition 2.2, the semiparametric efficiency bound for τ_1 is given by

$$V_1^{\star\star} = \mathbb{E}_P \left[\frac{\gamma(X)}{\pi^2} \left(\left(\frac{\gamma(X)}{\gamma(S,X)} \frac{1-\gamma(S,X)}{1-\gamma(X)} \frac{\varrho(S,X) - \varrho(X)}{\varrho(X)(1-\varrho(X))} \right)^2 \sigma^2(S,X) \right.\right. \\ \left.\left. + (\overline{\nu}_1(X) - \overline{\nu}_0(X) - \tau_1)^2 + \Lambda_{0,1}(S,X) + \Lambda_{1,1}(S,X) \right) \right] . \tag{3.13}$$

Remark 3.3. In Appendix C.2, we analyze how the semiparametric efficiency bounds derived in Corollary 3.1 change if different components of the nuisance parameters η or φ are known a priori. In both models, if the classical propensity score $\varrho(X)$, i.e., the probability of being assigned to treatment in the experimental sample as a function of covariates, is known, then the semiparametric efficiency bounds are unchanged.[8] This echoes an analogous ancillarity result for estimation of average treatment effects under unconfoundedness given in Hahn (1998).

By contrast, both semiparametric efficiency bounds change if the propensity score $\gamma(x)$, i.e., the probability of being assigned to the observational sample as a function of covariates, is known. This result is relevant for settings where the experimental sample is known to be drawn from the same population as the observational sample and indicates that development of estimators tailored to this setting may be fruitful.[9] In the Statistical Surrogacy Model, somewhat curiously, the efficient influence function $\xi_1(b, \tau_1, \varphi)$ is additionally invariant to knowledge of the distribution, in the observational sample, of the short-term outcomes conditional on covariates, i.e. the law

[8] Invariance of the semiparametric efficiency bounds to knowledge of the propensity score $\varrho(X)$ would no longer hold if the estimands of interest were average long-term effects for the treated population.

[9] On the other hand, in that context, consideration of the unconditional long-term treatment effect $\tau = \mathbb{E}[Y(1) - Y(0)]$ is tenable and natural. We expect the efficiency bound for this functional to be invariant to knowledge of the propensity score $\gamma(x)$.

$S \mid X, G = 1$. This invariance is a consequence of the choice (A.27) in construction of the efficient influence function $\xi_1(b, \tau_1, \varphi)$ in the proof of Theorem 3.1. ■

Next, we demonstrate that the efficient influence functions $\psi_1(b, \tau_1, \eta)$ and $\xi_1(b, \tau_1, \varphi)$ expressed in Theorem 3.1 are, in fact, the unique influence functions for τ_1 in their respective models. We recall that an influence for τ_1 is any mean-zero and square integrable function function $\tilde{\psi}(b)$ that satisfies the condition

$$\tau_1' = \mathbb{E}_P[\tilde{\psi}(B)\ell'(B)] ,$$

where τ_1' is the pathwise derivative of τ_1 along an arbitrary parametric submodel evaluated at zero and $\ell'(B)$ is the score function of this submodel; see e.g., Chapter 25 of Van der Vaart (1998) for further discussion.

Theorem 3.2. *There are unique influence functions in each model:*

(1) Under the Latent Unconfounded Treatment Model, given in Definition 2.1, $\psi_1(b, \tau_1, \eta)$ is the unique influence function for τ_1.

(2) Under the Statistical Surrogacy Model, given in Definition 2.2, $\xi_1(b, \tau_1, \varphi)$ is the unique influence function for τ_1.

Remark 3.4. Let $\mathcal{P} \subset \mathcal{M}_\lambda$ denote the set of probability distributions that satisfy either the Latent Unconfounded Treatment Model or the Statistical Surrogacy Model. In the terminology of Chen and Santos (2018), Theorem 3.2, Part (1), demonstrates that P is locally just-identified by $\mathcal{P}$. As a result, by Theorem 3.1 of Chen and Santos (2018), all regular and asymptotically linear (RAL) estimators of τ_1 are first-order equivalent under the maintained assumptions. In particular, there are no RAL estimators of τ_1 that have smaller asymptotic variances than others.

Consequently, semiparametric efficiency is equivalent to regularity and asymptotic linearity under the maintained assumptions. We note that if the propensity score $\varrho(x)$ admits known restrictions, then the resultant model would be semiparametrically over-identified. In this case, not all RAL estimators are first-order equivalent. However, since the efficiency bound does not change, the estimators we propose in the following section remain efficient with known propensity score.[10]

Moreover, again by Theorem 3.1 of Chen and Santos (2018), the model $\mathcal{P}$ does not have any locally testable restrictions in the sense that there are no specification tests of the maintained

[10] The semiparametric literature on average treatment effect and local average treatment effect estimation (e.g. Frölich, 2007; Hirano et al., 2003) discusses efficiency as well, despite the fact that the models in question are similarly just-identified.

identifying assumptions with nontrivial local asymptotic power. Analogous statements follow from Theorem 3.2, Part (2). In that sense, the maintained identifying assumptions are minimal. ■

4. ESTIMATION

We now consider estimation of the long-term average treatment effect τ_1.[11] The estimators that we consider can each be viewed as semiparametric Z-estimators associated with an identifying moment function. That is, each estimator is premised on determining the value τ_1 that solves a sample analogue of a moment condition

$$\mathbb{E}_P\left[g(B_i, \tau_1, \zeta)\right] = 0\ ,$$

for some identifying moment function $g(\cdot)$, where ζ is an unknown nuisance parameter, replaced with its estimated counterpart in practice.

Our treatment differs by whether the identifying moment function $g(\cdot)$ is given by the efficient influence functions $\psi_1(\cdot)$ or $\xi_1(\cdot)$, derived in Theorem 3.1 or given by some other moment condition. Moment conditions defined by influence functions are often referred to as Neyman orthogonal moment conditions. We adapt very general arguments from Chernozhukov et al. (2018) to establish that these estimators are consistent and asymptotically normal.

Our consideration of estimators based on non-orthogonal moments is selective and is more specialized. To obtain theoretical guarantees, we restrict attention to estimators that plug-in nuisance parameter estimates derived from the method of sieves (Chen and Liao, 2015; Chen et al., 2014). Sufficient conditions for estimators with this structure are generally available, but are more delicate and difficult to verify. We state and verify these conditions for one of the estimators that we consider.[12]

Throughout, it is useful to keep in mind that the efficient influence functions $\psi_1(\cdot)$ and $\xi_1(\cdot)$ are the only influence functions in their respective models (i.e., Theorem 3.2). Consequently, all regular and asymptotically linear estimators are first-order equivalent, that is, their asymptotic variances are all equal to the semiparametric efficiency bound. Thus, the asymptotic variances of different estimators are the same, but the conditions under which they are asymptotically normal may be different.

[11] In Appendix B, we provide an analogous treatment of estimators of the long-term average treatment effect τ_0 in the experimental population.

[12] The general high-level conditions in Chen and Liao (2015) and Chen et al. (2014) apply to each of estimators that we consider. Lower-level conditions, analogous to those discussed in Section 4.2 will exist for these estimators as well. However, as the derivation of these conditions is lengthy and cumbersome, we provide an illustration of this argument for only one estimator.

4.1 Orthogonal Moments. We begin by considering estimators that are built directly on the influence functions $\psi_1(\cdot)$ or $\xi_1(\cdot)$, derived in Theorem 3.1 with the "Double/Debiased Machine Learning" (DML) construction developed in Chernozhukov et al. (2018).

4.1.1 Construction. The DML construction proceeds in two steps. First, estimates of the nuisance functions η or φ, defined in Theorem 3.1, are computed with cross-fitting. Second, estimates of τ_1 are obtained by plugging the estimated values of η and φ into their respective efficient influence functions and solving for the values of τ_1 that equate the sample means of these estimates of the efficient influence functions with zero.

Definition 4.1 (DML Estimators)**.** Let $\hat{\eta}(I)$ and $\hat{\varphi}(I)$ denote generic estimates of η and φ based on the data $\{B_i\}_{i\in I}$ for some subset $I \subseteq [n]$. Let $\{I_\ell\}_{\ell=1}^k$ denote a random k-fold partition of $[n]$ such that the size of each fold is $m = n/k$. The estimator $\hat{\tau}_{1,\mathsf{DML}}$ is defined as the solution to

$$\frac{1}{k}\sum_{\ell=1}^{k}\frac{1}{m}\sum_{i\in I_\ell}\psi_1(B_i, \hat{\tau}_{1,\mathsf{DML}}, \hat{\eta}(I_\ell^c)) = 0 \quad \text{or} \quad \frac{1}{k}\sum_{\ell=1}^{k}\frac{1}{m}\sum_{i\in I_\ell}\xi_1(B_i, \hat{\tau}_{1,\mathsf{DML}}, \hat{\varphi}(I_\ell^c)) = 0$$

for the Latent Unconfounded Treatment and Statistical Surrogacy Models, respectively.

Remark 4.1. The fundamental structures underlying standard estimators of average treatment effects under unconfoundedness can be classified as being based on either "inverse propensity score weighting" or "outcome regression" (Imbens, 2004); elements of each structure appear in the estimators formulated in Definition 4.1.[13] Inverse propensity score weighted (IPW) estimators, also referred to as Horvitz and Thompson (1952) estimators, are constructed by weighting the observed values of outcomes by their inverse propensity scores; see e.g., Rosenbaum and Rubin (1983) and Hirano et al. (2003). By contrast, outcome regression estimators are constructed by imputing unobserved potential outcomes with estimates of their expectation conditioned on covariates.

The estimators formulated in Definition 4.1 combine IPW and outcome regression components with an error-correcting structure comparable to the augmented inverse propensity weighted (AIPW) estimator of Robins et al. (1995). To illustrate, observe that $\psi_1(b, \tau_1, \eta)$ can be interpreted as first approximating τ_1 with the outcome regression $\overline{\mu}_1(x) - \overline{\mu}_0(x)$ in the observational sample. Heuristically, the biases in this approximation, e.g., induced by regularization, are then corrected by applying IPW to the residuals of the approximation of $\overline{\mu}_w(x)$ to $\mu_w(s, x)$ with

$$\frac{w(\mu_1(s,x) - \overline{\mu}_1(x))}{\varrho(x)} - \frac{(1-w)(\mu_0(s,x) - \overline{\mu}_0(x))}{1-\varrho(x)}$$

[13] Imbens (2004) also discusses estimators based on matching and Bayesian calculations. We do not develop estimators with these structures in this paper, and view their consideration as an interesting extension.

computed in the experimental sample and reweighed by $\gamma(x)/(1-\gamma(x))$ to represent an expectation over the observational sample. However, the correction above may introduce additional biases through the estimation of $\mu_w(s,x)$; these are in turn corrected by applying IPW to the residuals of the approximation of $\mu_w(s,x)$ to $Y(w)$ with

$$\frac{w(y-\mu_1(s,x))}{\rho_1(s,x)} - \frac{(1-w)(y-\mu_0(s,x))}{\rho_0(s,x)}$$

computed in the experimental sample. An analogous interpretation can be formulated for the structure of the efficient influence function $\xi_1(b,\tau_1,\varphi)$.[14] Further discussion, at varying levels of rigor, of this "bias-correction" interpretation of the structure of estimators based on efficient influence functions is given in Section 4 of Kennedy (2023) and Chapter 7 of Bickel et al. (1993).■

Remark 4.2. At a high-level, the cross-fitting construction used in Definition 4.1 is implemented so that the estimation errors, e.g., $\mu_w(S_i,X_i) - \hat{\mu}_w(S_i,X_i)$, and model errors, e.g., $Y(W_i) - \mu_w(S_i,X_i)$, are unrelated for a given observation. Association between these two forms of error may have particularly pernicious effects in finite-samples if estimates of nuisance functions suffer from over-fitting. More technically, cross-fitting allows us to avoid imposing Donsker-type regularity conditions in our asymptotic analysis, which would exclude estimators with non-negligible asymptotic regularization. Standard implementations of popular machine learning algorithms may feature such regularization; see Chernozhukov et al. (2016) for detailed discussion and illustration of this point. Further discussion of cross-fitting methods in semiparametric estimation is given in Klaassen (1987) and Newey and Robins (2018). ■

Remark 4.3. We require specialized approaches to estimate the nuisance functions $\rho_w(s,x)$, $\overline{\mu}_w(x)$, and $\overline{\nu}_w(x)$. We estimate $\rho_w(s,x)$ by combining separate estimates of each of the objects displayed in Equation (3.9). We estimate $\overline{\mu}_w(x)$ and $\overline{\nu}_w(x)$ by first computing estimates of $\mu_w(s,x)$ and $\nu(s,x)$ in the observational sample, denoted by $\hat{\mu}_w(s,x)$ and $\hat{\nu}(s,x)$, and then computing estimates of $\mathbb{E}_P[\hat{\mu}_w(s,x)|W=w,X=x,G=0]$ and $\mathbb{E}_P[\hat{\nu}(s,x)|W=w,X=x,G=0]$ in the experimental sample. In Appendix C.3, we derive the rate of convergence for particular implementations of estimators with this structure based on linear sieves. ■

4.1.2 Large-Sample Theory. We now study the asymptotic behavior of the estimators formulated in Definition 4.1. First, in Appendix C.1, we demonstrate that, under weak regularity conditions and under both models, the estimator $\hat{\tau}_{1,\mathsf{DML}}$ is consistent for τ_1 if either the long-term outcome means

[14] Note that in the term corresponding to the observational sample in $\xi_1(\cdot)$, the unobserved treatment indictor w is replaced by the probability of treatment conditional on short-term outcomes and covariates.

or the propensity scores are estimated consistently. Second, we establish asymptotic normality by providing conditions sufficient for the application of Theorem 3.1 of Chernozhukov et al. (2018). We impose a set of standard bounds on moments of the data, and a set of conditions on the uniform rates of convergence of nuisance parameter estimators. Throughout, for a collection of scalar-valued nuisance parameters $\theta = (\theta_1, \dots, \theta_\ell)$, we let $\|\theta\|_{P,q} = \max_{i\in[\ell]} \left(\mathbb{E}_P|\theta_i(B)|^q\right)^{1/q}$.

Assumption 4.1 (Moment Bounds). *Let $C, c > 0$ be constants. Under the Latent Unconfounded Treatment Model, the moment bounds*

$$\|Y(w)\|_{P,q} \leq C, \quad \mathbb{E}_P\left[\sigma_w^2(S,X) \mid X\right] \leq C,$$
$$\mathbb{E}_P\left[(Y(w) - \mu_w(S,X))^2\right] \geq c, \quad \text{and} \quad c \leq \mathbb{E}_P\left[(\mu_w(S,X) - \overline{\mu}_w(x))^2 \mid X\right] \leq C$$

hold for each $w \in \{0,1\}$ and λ-almost every X. Analogous bounds hold for the Statistical Surrogacy Model, where $\sigma(S,X)$, $\nu(S,X)$, and $\overline{\nu}_w(x)$ replace $\sigma_w(S,X)$, $\mu_w(S,X)$, and $\overline{\mu}_w(x)$, respectively.

Assumption 4.2 (Convergence Rates). *Let $\mathcal{P} \subset \mathcal{M}_\lambda$ be the set of all probability distributions P that satisfy the Latent Unconfounded Treatment Model stated in Definition 2.1. Consider a sequence of estimators $\hat{\eta}_n(I_n^c) = (\hat{\omega}_{\psi,n}, \hat{\kappa}_{\psi,n}, \hat{\pi}_n)$ indexed by n, where $I_n \subset [n]$ is a random subset of size $m = n/k$ and $\hat{\pi}_n = \frac{1}{n-m}\sum_{i\in I_n^c} G_i$. For some sequences $\Delta_n \to 0$ and $\delta_n \to 0$ and constants $\varepsilon, C > 0$ and $q > 2$, that do not depend on P, with P-probability at least $1 - \Delta_n$,*

1. (Consistency in 2-norm) $\quad n^{-1/2} \leq \|\hat{\eta}_n - \eta\|_{P,2} \leq \delta_n$,

2. (Boundedness in q-norm) $\quad \|\hat{\eta}_n - \eta\|_{P,q} < C$,

3. (Non-degeneracy) $\quad \varepsilon \leq \hat{\kappa}_n \leq 1 - \varepsilon$, where the inequalities apply entry-wise, and

4. ($o(n^{-1/2})$ product rates) $\quad \|\hat{\omega}_{\psi,n} - \omega_\psi\|_{P,2} \cdot \|\hat{\kappa}_{\psi,n} - \kappa_\psi\|_{P,2} \leq \delta_n n^{-1/2}$.

Analogous conditions hold for the Statistical Surrogacy Model, where φ replaces η.

Remark 4.4. Assumption 4.2 imposes the restriction that the product of the estimation errors for the long-term outcome means and propensity scores converges at the rate $o(n^{-1/2})$.[15] These rates can be achieved, even if the dimensionality of the covariates or the short-term outcomes is increasing with n, by many standard machine learning algorithms including the Lasso and Dantzig selector (Belloni et al., 2014; Bickel et al., 2009), boosting algorithms (Luo and Spindler, 2017), regression trees and random forests (Wager and Walther, 2015), and neural networks (Chen and White, 1999; Farrell et al., 2021) under appropriate conditions on the structure or sparsity of the underlying model.

[15] If the true propensity score $\varrho(x)$ is known, then using this information when constructing $\hat{\tau}_{1,\mathsf{DML}}$ may produce an estimator that performs well in finite-samples. However, plugging the true propensity score into an estimator based off of a non-orthogonal moment would probably be inefficient. For example, in Hahn (1998), an IPW type estimator for the average treatment effect based off of the true propensity score is shown to be inefficient.

In Appendix C.3, we verify that estimates of $\overline{\mu}_w(x)$ and $\overline{\nu}_w(x)$ based on linear sieves can achieve these rates under sufficiently stringent restrictions on the smoothness of the long-term outcome means. It is reasonable to expect that analogous results should be available for more complicated estimators, e.g., featuring penalization or more complicated bases. Results of this form are an interesting direction for further research. ■

Theorem 4.1 establishes the asymptotic properties of the estimators formulated in Definition 4.1. For the sake of brevity, we state the result only for the Latent Unconfounded Treatment Model. An analogous result holds for the Statistical Surrogacy Model. We provide proofs for both results in Appendix A.7.

Theorem 4.1. *Let $\mathcal{P} \subset \mathcal{M}_\lambda$ be the set of all probability distributions P for $\{B_i\}_{i=1}^n$ that satisfy the Latent Unconfounded Treatment Model stated in Definition 2.1 in addition to Assumption 4.1. If Assumption 4.2 holds for $\mathcal{P}$, then*

$$\sqrt{n}(\hat{\tau}_{1,\mathsf{DML}} - \tau_1) \xrightarrow{d} \mathcal{N}(0, V_1^\star) \tag{4.1}$$

uniformly over $P \in \mathcal{P}$, where $\hat{\tau}_{1,\mathsf{DML}}$ is defined in Definition 4.1, $V_1^\star$ is defined in Corollary 3.1, and $\xrightarrow{d}$ denotes convergence in distribution. Moreover, we have that

$$\hat{V}_1^\star = \frac{1}{k}\sum_{\ell=1}^{k} \frac{1}{m}\sum_{i \in I_\ell} (\psi_1(B_i, \hat{\tau}_{1,\mathsf{DML}}, \hat{\eta}(I_\ell^c)))^2 \xrightarrow{p} V_1^\star \tag{4.2}$$

uniformly over $P \in \mathcal{P}$, where $\xrightarrow{p}$ denotes convergence in probability. As a result, we obtain the uniform asymptotic validity of the confidence intervals

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left| P\left(\tau_1 \in \left[\hat{\tau}_{1,\mathsf{DML}} \pm z_{1-\alpha/2}\sqrt{\hat{V}_1^\star/n}\right]\right) - (1-\alpha)\right| = 0\,, \tag{4.3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

4.2 Non-orthogonal Moments. The moment functions $\psi_1(\cdot)$ and $\xi_1(\cdot)$ considered in Section 4.1 are quite complicated. Constructing the estimator $\hat{\tau}_{1,\mathsf{DML}}$ requires estimating several propensity scores and long-term outcome means. It is natural to ask whether it suffices to consider simpler estimators based on moment conditions with fewer nuisance functions.

In this section, we consider a suite of estimators that do not use the efficient influence functions $\psi_1(\cdot)$ or $\xi_1(\cdot)$ as identifying moment functions. We emphasize estimators that bear a similarity to standard IPW or outcome regression estimators for average treatment effects, several of which were initially proposed in Athey et al. (2020a) and Athey et al. (2020b). We verify the asymptotic normality of one of these estimators, when nuisance parameters are estimated with the method of sieves, following ideas developed in (Chen and Liao, 2015; Chen et al., 2014).

4.2.1 Construction. Each of the estimators that we consider can be viewed as a Z-estimator based on a moment function $g(\cdot, \tau_1, \zeta)$, taking as an argument an unknown nuisance parameter ζ.

Definition 4.2 (Non-orthogonal Moment Estimators)**.** The estimator $\hat{\tau}_1(g)$ is defined as the solution to the sample moment condition

$$\frac{1}{n}\sum_{i=1}^{n} g(B_i, \hat{\tau}_1(g), \hat{\zeta}) = 0 \ ,$$

where $\hat{\zeta}$ is a generic estimate of ζ based on the data $\{B_i\}_{i=1}^n$.

We consider two classes of moment functions, differing in whether or not the estimators that they entail resemble IPW or outcome regression estimators for average treatment effects. In the Latent Unconfounded Treatment Model, the random variable

$$g_{\mathsf{w}}(B_i, \tau_1, \zeta_{\mathsf{w}}) = \frac{G_i}{\pi}\left(\frac{W_i Y_i}{\rho_1(S_i, X_i)} - \frac{(1-W_i)Y_i}{\rho_0(S_i, X_i)}\right) - \tau_1 \ , \tag{4.4}$$

where $\zeta_{\mathsf{w}} = (\pi, \rho_1, \rho_0)$, has mean zero under P. The estimator $\hat{\tau}_1(g_{\mathsf{w}})$, proposed originally by Athey et al. (2020a), only requires estimation of the nuisance parameters in ζ_{w} and can be viewed as an analogue to the IPW estimator for average treatment effects. In turn, the functions

$$g_{\mathsf{or},1}(B_i, \tau_1, \zeta_{\mathsf{or},1}) = \frac{G_i}{\pi}\left(\overline{\mu}_1(X_i) - \overline{\mu}_0(X_i)\right) \quad \text{and} \tag{4.5}$$

$$g_{\mathsf{or},0}(B_i, \tau_1, \zeta_{\mathsf{or},0}) = \frac{1-G_i}{\pi}\frac{\gamma(X_i)}{1-\gamma(X_i)}\left(\mu_1(S_i, X_i) - \mu_0(S_i, X_i)\right) \ , \tag{4.6}$$

where $\zeta_{\mathsf{or},1}$ and $\zeta_{\mathsf{or},0}$ collect nuisance parameters, yield the outcome regression type estimators $\hat{\tau}_1(g_{\mathsf{or},1})$ and $\hat{\tau}_1(g_{\mathsf{or},0})$. The estimator $\hat{\tau}_1(g_{\mathsf{or},0})$ was originally proposed by Athey et al. (2020a).

Analogously, in the Statistical Surrogacy Model, the moment function

$$h_{\mathsf{w}}(B_i, \tau_1, \varsigma_{\mathsf{w}}) = \frac{G_i Y_i}{\pi}\frac{\gamma(X_i)}{1-\gamma(X_i)}\frac{1-\gamma(S_i, X_i)}{\gamma(S_i, X_i)}\left(\frac{\varrho(S_i, X_i)}{\varrho(X_i)} - \frac{1-\varrho(S_i, X_i)}{1-\varrho(X_i)}\right) - \tau_1 \ , \tag{4.7}$$

where ς_{w} collects nuisance parameters, yields an IPW-type estimator $\hat{\tau}_1(h_{\mathsf{w}})$ that is similar to an estimator proposed by Athey et al. (2020b). The moment functions

$$h_{\mathsf{or},1}(B_i, \tau_1, \varsigma_{\mathsf{or},1}) = \frac{G_i}{\pi}\left(\overline{\nu}_1(X_i) - \overline{\nu}_0(X_i)\right) \quad \text{and} \tag{4.8}$$

$$h_{\mathsf{or},0}(B_i, \tau_1, \varsigma_{\mathsf{or},0}) = \frac{1-G_i}{\pi}\frac{\gamma(X_i)}{1-\gamma(X_i)}\left(\frac{W_i}{\varrho(X_i)} - \frac{1-W_i}{1-\varrho(X_i)}\right)\nu(S_i, X_i) \ , \tag{4.9}$$

result in the outcome regression type estimators $\hat{\tau}_1(h_{\mathsf{or},1})$ and $\hat{\tau}_1(h_{\mathsf{or},0})$, respectively. The estimator $\hat{\tau}_1(h_{\mathsf{or},0})$ was originally proposed by Athey et al. (2020b). Dynarski et al. (2021) use an estimator closely related to $\hat{\tau}_1(h_{\mathsf{or},0})$ with an estimate of $\nu(s,x)$ based on linear regression, in their analysis of the effects of college tuition grants on college complteion rates.

4.2.2 Large-Sample Theory. Theoretical analysis of the large-sample performance of estimators based on non-orthogonal moment conditions requires a more specialized treatment. Sufficient conditions for their asymptotic normality in the literature are often more delicate and stronger than those for the DML estimators considered in Section 4.1. We provide details of this analysis for the estimator $\hat{\tau}_1(g_{\mathsf{w}})$ only, as stating and verifying sufficient conditions for asymptotic linearity is cumbersome. Nevertheless, the basic structure of the conditions that we pose, and the method of their verification, is applicable to each of the estimators formulated above.

Constructing the estimator $\hat{\tau}_1(g_{\mathsf{w}})$ requires an estimate of the nuisance parameter ζ_{w}. We restrict attention to procedures that estimate the remaining components of ζ_{w}, i.e., ρ_1 and ρ_0, with linear sieves. We detail this procedure in Appendix A.8.

Theorem 4.2 establishes the asymptotic normality of the estimator $\hat{\tau}_1(g_{\mathsf{w}})$ when the nuisance parameter ζ_{w} is estimated with the method of sieves. Stating the precise sufficient conditions for this Theorem requires additional notation and definitions, which we defer to Appendix A.8.

Theorem 4.2. *Under the Latent Unconfounded Treatment Model and Assumption A.3 stated in Appendix A.8. we have that*

$$\sqrt{n}(\hat{\tau}_1(g_{\mathsf{w}}) - \tau_1) \xrightarrow{d} \mathcal{N}(0, V_1^\star) \, ,$$

where $V_1^\star$ is the semiparametric efficiency bound for τ_1 defined in Corollary 3.1.

We again emphasize that, by Theorem 3.2, *any* regular and asymptotically linear estimator $\hat{\tau}_1$ for τ_1 achieves the semiparametric efficiency bound $V_1^\star$, if the assumptions that define the Latent Unconfounded Treatment Model are the *only* set of restrictions that are imposed. Thus, the semiparametric efficiency of $\hat{\tau}_1(g_{\mathsf{w}})$, per se, is to be expected. The substance of Theorem 4.2 is the asymptotic linearity.

Remark 4.5. Informally, Assumption A.3 requires that:

(1) $\hat{\zeta}_{\mathsf{w}}$ is $o(n^{-1/4})$-consistent;

(2) The parameter space containing ζ_{w} is Donsker; and

(3) The sieve space chosen to approximate ζ_{w} has limited complexity.

Condition (1) is analogous to the product rate condition in Assumption 4.2. It is weaker in the sense that no consistent estimators for the nuisance parameters in η that are not in ζ_{w} are needed. On the other hand, Condition (1) places more stringent conditions on the rate that $\hat{\zeta}_{\mathsf{w}}$ estimates ζ_{w}. In particular, Assumption 4.2 will still hold in situations where some elements of ζ_{w} are estimated at a rate slower than $o(n^{-1/4})$, so long as the product of the errors in estimation of the long-term outcome means and propensity scores is smaller than $o(n^{-1/2})$.[16] Condition (2) is imposed in order to ensure stochastic equicontinuity for the moment condition $\zeta_{\mathsf{w}} \mapsto g_{\mathsf{w}}(\cdot, \tau_1, \zeta_{\mathsf{w}})$ treated as a process indexed by ζ_{w}. This condition ensures that estimating τ_1 and ζ_{w} using the same data does not induce errors that are excessively large. Using sample-splitting would eliminate the need for this condition. Condition (3) is specific to the sieve approach for estimating nuisance parameters. It is not directly imposed in Assumption 4.2, but may be needed to justify rate conditions when one estimates nuisance parameters with sieves. ■

5. SIMULATION

We now compare the estimators formulated in Section 4 with a simulation calibrated to data from Banerjee et al. (2015). We find that the DML estimators considered in Section 4.1 are more accurate than the estimators based on non-orthogonal moments considered in Section 4.2, particularly if a nonparametric approach is taken to nuisance parameter estimation.

5.1 Data, Calibration, and Design. Banerjee et al. (2015) study randomized evaluations of several similar poverty-alleviation programs implemented by BRAC, a large non-governmental organization. These programs allocated productive assets (typically livestock) to participating households and measured both short-term and long-term economic outcomes.

We restrict our attention to data from the evaluation of the program implemented in Pakistan. For each of the 854 households in our cleaned sample, survey measurements of the consumption levels, food security, assets, savings, and outstanding loans were taken prior to, as well as two and three years after, treatment. We use the pre-treatment measurements as covariates (i.e., X_i), the two-year post-treatment measurements as short-term outcomes (i.e., S_i), and the three-year post-treatment measurements as long-term outcomes (i.e., Y_i). There are 20 pre-treatment covariates and 21 short-term outcomes. In the main text, the long-term outcome of interest is total household assets; we give analogous results for total household consumption in Appendix D.5. Appendix D.1 gives further information on the construction and content of these data.

We calibrate a generative model to these data with a Generative Adversarial Network (Goodfellow et al., 2014), following a method for simulation design developed in Athey et al. (2021). The details

[16] Imbens (2004) outlines similar heuristics for average treatment effect estimation under unconfoundedness.

of this calibration are given in Appendix D.2.[17] Crucially, a sample drawn from this model consists of covariates and both treated and untreated short-term and long-term potential outcomes for a hypothetical household (the vector $(X_i, S_i(1), S_i(0), Y_i(1), Y_i(0))$ in our notation). That is, we observe the true short-term and long-term treatment effects for each household sampled from this model, and can measure true long-term average treatment effects by averaging over many simulation draws.

With this calibrated model, we generate a collection of hypothetical data sets that satisfy either the Latent Unconfounded Treatment Model or the Statistical Surrogacy Model, as desired. The quality of various estimators is then determined by measuring their average accuracy in recovering long-term average treatment effects in a variety of metrics. To generate a hypothetical data set, we draw a collection of samples from the generative mode of size $h \cdot n$, where h is some multiplier that we vary and n is the sample size of the Banerjee et al. (2015) data. Each observation is assigned to being either "experimental" or "observational" with probability $1/2$, and so the experimental and observational samples have identical distribution of covariates. Treatment for experimental samples is always assigned uniformly at random. Treatment for observational samples is assigned with possible confounding. Specifically, we determine treatment probabilities with an increasing function, indexed by a parameter ϕ, of each hypothetical household's true short-term treatment effects. Larger values of ϕ indicate more confounding. Details of this simulation design and parameterization of confounding are given in Appendix D.3.

5.2 Comparison of Methods. We begin by comparing the estimators formulated in Section 4 with a simple difference between the mean long-term treated and untreated outcomes in the observational sample

$$\hat{\tau}_{1,DM} = \frac{\sum_i G_i W_i Y_i}{\sum_i G_i W_i} - \frac{\sum_i G_i (1 - W_i) Y_i}{\sum_i G_i (1 - W_i)} \,. \tag{5.1}$$

This estimator is naive, making no adjustment for confounding in the observational sample, and is infeasible if treatment is not observed in the observational sample.

Figure 2 compares the absolute bias and root mean squared error of the estimators formulated in Section 4 with the naive estimator (5.1).[18] We compare the use of the Lasso (Tibshirani, 1996), Generalized Random Forests (Athey et al., 2019), and XGBoost (Chen and Guestrin, 2016) for nuisance parameter estimation. Details about the implementation of these estimators are given in

[17] In Appendix D.2, we demonstrate that the joint distribution of data drawn from this model matches the leading moments of joint distribution of the data from Banerjee et al. (2015) remarkably closely.

[18] Measurements of the variance of each estimator are displayed in Figures D.9 and D.10. We note that in the Latent Unconfounded Treatment Model, one of the outcome regression estimators has very small variance when nuisance parameters are estimated with linear regression. However, the bias of this estimator is very large.

Appendix D.4. Panels A and B display results for the Latent Unconfounded Treatment and Statistical Surrogacy Models, respectively. Columns within each panel vary the sample size multiplier h. Each column displays results for the confounding parameter ϕ set to zero, indicating no confounding in the observational sample and labeled as "Baseline," in addition to two parameterizations indicating non-zero confounding.

In both the Latent Unconfounded Treatment Model and the Statistical Surrogacy Model, the biases of the DML estimators considered in Section 4.1 tend to be substantially smaller than the biases of the alternative outcome regression or weighting estimators considered in Section 4.2, so long as the nuisance parameters are not estimated by linear regression. The weighting estimator has performance more comparable to the DML estimator in the Latent Unconfounded Treatment Model. The performances of the weighting estimator and the outcome regression estimators in the Statistical Surrogacy Model are more similar.

Figure 3 displays measurements of estimator quality for just the DML estimators in a format analogous to Figure 2, with the addition of a third row measuring one minus the coverage probability of confidence intervals constructed around each estimator. We report these estimates of coverage probabilities in Supplementary Appendix D.5. Confidence intervals constructed with the intervals given in (A.42) around the estimators formulated in Definition 4.1 have coverage probabilities that are reasonably close to the nominal level.

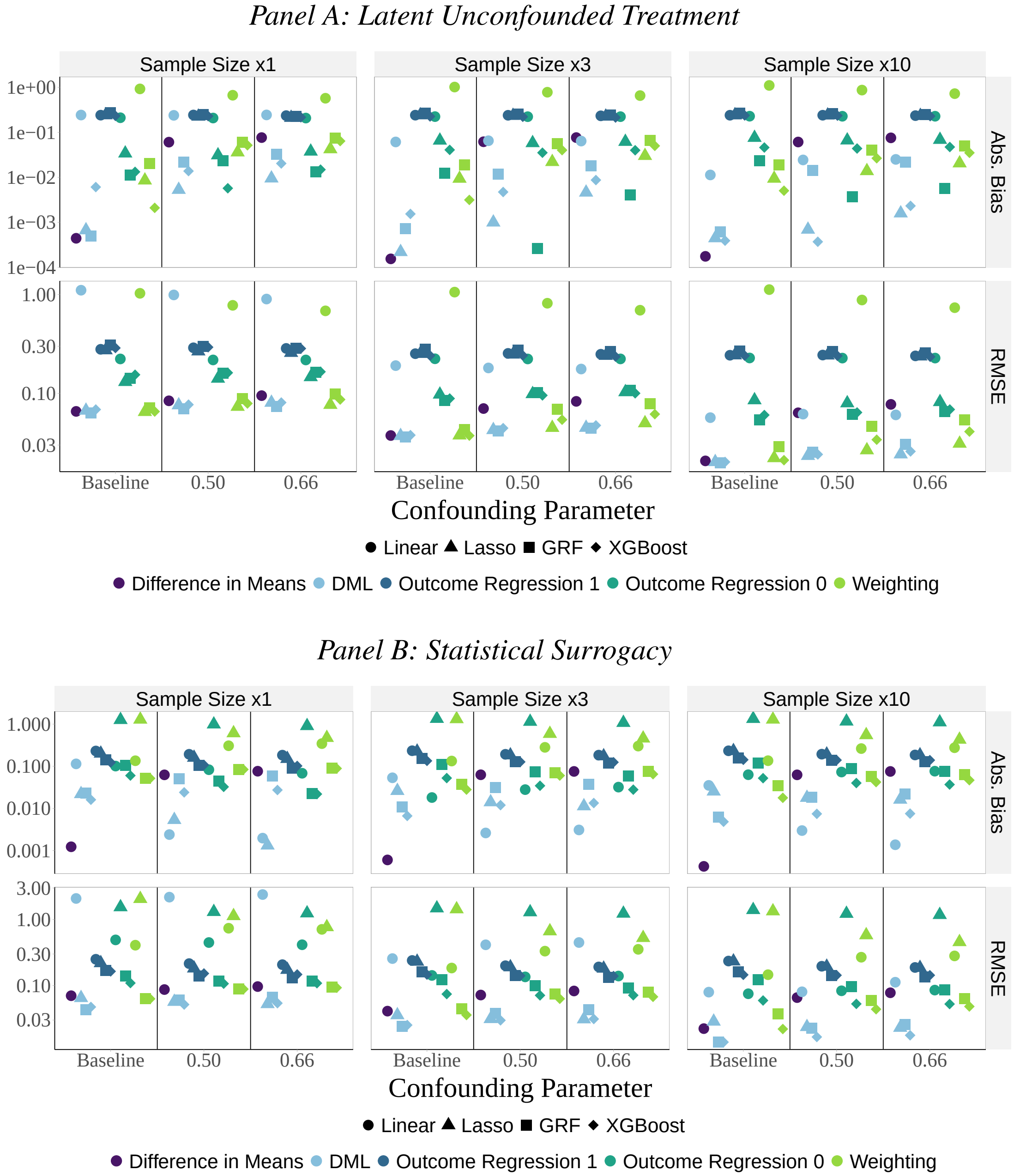
6. CONCLUSION

We study the estimation of long-term treatment effects through the combination of short-term experimental and long-term observational data sets. We derive efficient influence functions and calculate corresponding semiparametric efficiency bounds for this problem. These calculations facilitate the development of estimators that accommodate the applications of standard machine learning algorithms for estimating nuisance parameters. We demonstrate with simulation that these estimators are able to recover long-term treatment effects in realistic settings.

Important unresolved practical issues remain. Methods for choosing valid and informative short-term outcomes and assessing of the sensitivity of estimates to violations of identifying assumptions would be valuable. Additional useful extensions include the incorporation of instruments and continuous treatments and the accommodation of settings with limited covariate overlap.

In the case of estimation of average treatment effects under unconfoundedness, recent promising work (e.g., Athey et al., 2018; Bradic et al., 2019; Tan, 2020) has developed estimators that are specifically optimized to handle high-dimensional covariates and are able to attain various notions

FIGURE 2. Comparison of Estimators



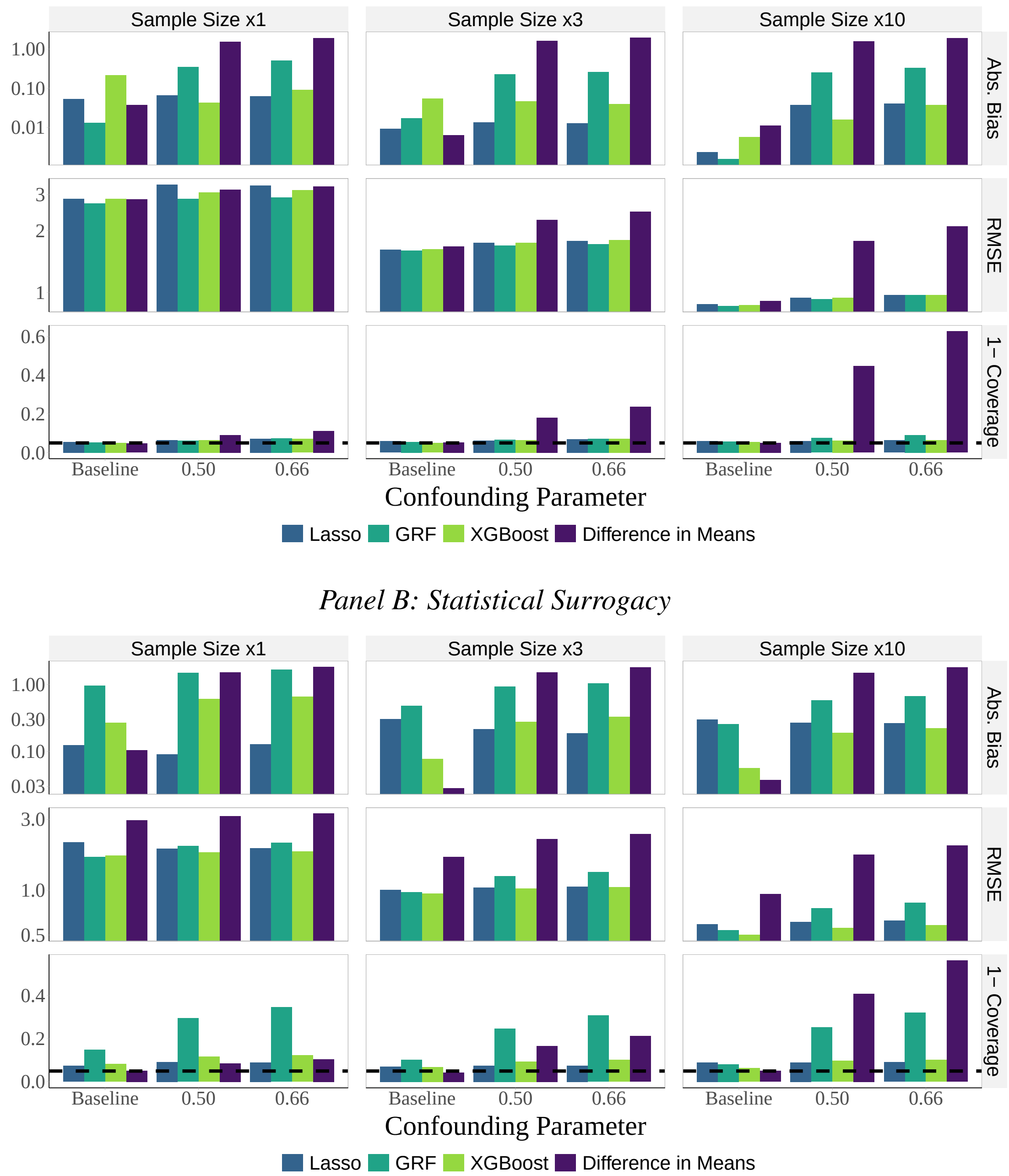
Panel A: Latent Unconfounded Treatment

Panel B: Statistical Surrogacy

Notes: Figure 2 compares measurements of the absolute bias and root mean squared error for the estimators formulated in Section 4, in addition to the difference in means estimator defined in (5.1). The y-axes are displayed in logs, base 10. The long-term outcome is total household assets. Panels A and B display results for the Latent Unconfounded Treatment and Statistical Surrogacy Models defined in Definition 2.1 and Definition 2.2, respectively. The columns of each panel vary the sample size multiplier h. Each sub-panel displays results for the baseline, unconfounded, case, as well as for the cases that the confounding parameter ϕ has been set to $1/2$ and $2/3$. Results for each estimator are displayed with dots of different colors. Results for different nuisance parameter estimators are displayed with dots of different shapes.

FIGURE 3. Finite-Sample Performance with Different Nuisance Parameter Estimators

Panel A: Latent Unconfounded Treatment



Panel B: Statistical Surrogacy

Notes: Figure 3 displays measurements of the quality of the estimators formulated in Section 4.1 implemented with several alternative choices of nuisance parameter estimators. The long-term outcome is total household assets. Panels A and B display results for the estimators defined in Definition 4.1 for the Latent Unconfounded Treatment and Statistical Surrogacy Models defined in Definition 2.1 and Definition 2.2, respectively. The columns of each panel vary the sample size multiplier h. The rows of each panel display the absolute value of the bias, one minus the coverage probability, and the root mean squared error of each estimator, from top to bottom, respectively. A dotted line denoting one minus the nominal coverage probability, 0.05, is displayed in each sub-panel in the third row. Each sub-panel displays a bar graph comparing measurements of the performance of the estimator defined in Definition 4.1, constructed with three types of nuisance parameter estimators, with the difference in means estimator (5.1). Each sub-panel displays results for the baseline, unconfounded, case, as well as for the cases that the confounding parameter ϕ has been set to $1/2$ and $2/3$.

of optimality under weak assumptions on, e.g., the sparsity of the outcome regression or propensity score models. It is not immediately clear how to apply these ideas to long-term average treatment effects. Further consideration of this problem would be a potentially valuable extension, as the resultant estimators may be particularly well-suited to the case where there are many short-term outcomes. Some progress on a related problem (where, effectively, there is a single short-term outcome) has been made by Viviano and Bradic (2021).

REFERENCES

Athey, S., Chetty, R., and Imbens, G. (2020a). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S., Chetty, R., Imbens, G., and Kang, H. (2020b). Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236).

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Begg, C. B. and Leung, D. H. (2000). On the use of surrogate end points in randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):15–28.

Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757 – 788.

Bia, M., Huber, M., and Lafférs, L. (2020). Double machine learning for sample selection models. *arXiv preprint arXiv:2012.00745*.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, pages 647–671.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Bouguen, A., Huang, Y., Kremer, M., and Miguel, E. (2019). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*.

Bradic, J., Wager, S., and Zhu, Y. (2019). Sparsity double robust inference of average treatment effects.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808 – 843.

Chen, X. and Liao, Z. (2015). Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186.

Chen, X., Liao, Z., and Sun, Y. (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178:639–658.

Chen, X. and Santos, A. (2018). Overidentification in regular models. *Econometrica*, 86(5):1771–1817.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Dal Bó, E., Finan, F., Li, N. Y., and Schechter, L. (2021). Information technology and government decentralization: Experimental evidence from paraguay. *Econometrica*, 89(2):677–701.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962.

Dynarski, S., Libassi, C., Michelmore, K., and Owen, S. (2021). Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students. *American Economic Review*, 111(6):1721–56.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Fisher, R. A. (1925). Statistical methods for research workers. *Statistical methods for research workers.*, (1st. ed).

Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75.

Gechter, M. and Meager, R. (2021). Combining experimental and observational studies in meta-analysis: A mutual debiasing approach.". Technical report, Mimeo.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S. M., and Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in jamaica. *Science*, 344(6187):998–1001.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2):437–452.

Gui, G. (2020). Combining observational and experimental data using first-stage covariates. *arXiv preprint arXiv:2010.05117*.

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., Curtis, M., Deng, A., Duan, W., Forbes, P., Frasca, B., Guy, T., Imbens, G. W., Saint Jacques, G., Kantawala, P., Katsev, I., Katzwer, M., Konutgan, M., Kunakova, E., Lee, M., Lee, M., Liu, J., McQueen, J., Najmi, A., Smith, B., Trehan, V., Vermeer, L., Walker, T., Wong, J., and Yashkov, I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieved June*, 9(2016):1–30.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Hotz, V. J., Imbens, G. W., and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270.

Hou, J., Mukherjee, R., and Cai, T. (2021). Efficient and robust semi-supervised estimation of ate with partially annotated treatment and response.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. (2022). Long-term causal inference under persistent confounding via data combination.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Kallus, N. and Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages

523–539.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: a review.

Klaassen, C. A. J. (1987). Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators. *The Annals of Statistics*, 15(4):1548 – 1562.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 129–157. University of California Press.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Luo, Y. and Spindler, M. (2017). ℓ_2-boosting for economic applications. *American Economic Review*, 107(5):270–73.

Muris, C. (2020). Efficient gmm estimation with incomplete data. *Review of Economics and Statistics*, 102(3):518–530.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. *Handbook of econometrics*, 6:5469–5547.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Rosenbaum, P. R. and Rubin, D. B. (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., and Baiocchi, M. (2020). Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Singh, R. (2021). A finite sample theorem for longitudinal causal inference with machine learning: Long term, dynamic, and mediated effects.

Singh, R. (2022). Generalized kernel ridge regression for long term causal inference: Treatment effects, dose responses, and counterfactual distributions.

Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837.

Tanner, J. C., Candland, T., and Odden, W. S. (2015). Later impacts of early childhood interventions: a systematic review. *Washington: Independent Evaluation Group, World Bank Group*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Twitter Engineering (2021). "We built a causal estimation framework on the idea of statistical 'surrogacy' (Athey et al 2016) - when we can't wait to observe long-run outcomes, we create a model based on intermediate data.". *Tweet*, @TwitterEng, 18 Oct, 12:18 P.M.

van der Laan, M. J. and Petersen, M. L. (2004). Estimation of direct and indirect causal effects in longitudinal studies.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Viviano, D. and Bradic, J. (2021). Dynamic covariate balancing: estimating treatment effects over time. *arXiv preprint arXiv:2103.01280*.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Yang, J., Eckles, D., Dhillon, P., and Aral, S. (2020). Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*.