The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely*

Susan Athey† Raj Chetty‡ Guido W. Imbens§ Hyunseung Kang¶

August 27, 2024

Abstract

A common challenge in estimating the impact of interventions (*e.g.*, job training programs, educational programs) is that many outcomes of interest (*e.g.*, lifetime earnings or other labor market outcomes) are observed with a long delay. In biomedical settings this is often addressed by using short-term outcomes as so-called "surrogates" for the outcome of interest, *e.g.*, tumor size as a surrogate for mortality in cancer studies. We build on this literature by combining multiple, possibly qualitatively distinct, short-term outcomes (*e.g.*, short-run earnings and employment indicators) systematically into a "surrogate index." Under the Prentice surrogacy assumption, which requires that the primary outcome is independent of the treatment conditional on the surrogates, we show that the average treatment effect on the surrogate index equals the treatment effect on the long-term outcome. We also relate the surrogacy assumption to a set of structural, causal assumptions. We then characterize the bias that arises from violations of each of the key assumptions, and we provide simple methods to validate these assumptions using additional observed outcomes. We apply our method to analyze the long-term impacts of a multi-site job training experiment in California. Rather than waiting a full nine years to directly observe the long-term impact, we show that it is possible to use short-term (the first six quarters) outcomes as surrogates. One could have estimated the program's long-term impacts on mean employment rates using the employment rates observed in the first six quarters, with a 35% reduction in standard errors.

Keywords: Potential Outcomes, Causality, Surrogate Outcomes, Surrogate Scpore, Surrogate Index, Mediators, Propensity Score, Principal Stratification, Job Training

*We are grateful for discussions with James Dailey, Lawrence Katz, David Ritzwoller, Dylan Small, Scott Stern, and Liang Xu and for comments from numerous seminar participants. We thank Kevin Chen, Yechan Park, Emanuel Schertz and James Stratton for outstanding research assistance. We are particularly grateful to Kevin Chen and David Ritzwoller for pointing out a mistake in the previous version of this paper. This research was funded through National Science Foundation Grant DMS-1502437, the Chan-Zuckerberg Initiative, the Bill & Melinda Gates Foundation, the Overdeck Foundation, ONR grants N00014-17-1-2131 and N00014-19-1-2468 and the Sloan Foundation.

†Graduate School of Business, Stanford University, and NBER, athey@stanford.edu.

‡Department of Economics, Harvard University, and NBER, chetty@fas.harvard.edu.

§Graduate School of Business, and Department of Economics, Stanford University, and NBER, imbens@stanford.edu.

¶Department of Statistics, University of Wisconsin at Madison, hyunseung@stat.wisc.edu.

1 Introduction

A fundamental challenge for evaluating interventions is that the primary outcomes of interest are often hard to measure. For example, researchers are often interested in the effect of the policy on some long-term outcome but do not observe that in their study. Instead, they observe a number of short-term outcomes that are all related to this primary outcome of interest.

One setting where this type of problem arises involves an educational policy maker evaluating a policy that would change class size. The ultimate goal may be to improve long-term labor market outcomes for the students. However, the decision regarding the class size policy needs to be made at a time when only short-term outcomes such as test scores or other educational achievement measures are available. Another setting involves policy makers considering labor market interventions such as job search assistance or human capital acquisition programs, where they may be primarily interested in the long-term labor market attachment of the participants, but in the short run they may only have access to outcomes such as employment records or earnings over a short period of time. In randomized experiments for medical interventions, the ultimate outcome of interest is often survival or quality-adjusted years of life. Survival rates may be high in the short run, and so typically such trials are evaluated in terms of surrogate measures, such as including tumor size or other measures of the progression of the disease, which can be measured earlier. In all of these types of setting, to make a timely decision, the policy maker needs to assess the programs based on short-term outcomes. These challenges also arise in business settings. In the context of experimentation in digital technology companies, a discussion of the most important challenges ranks as the top concern that "While most experiments in the industry run for 2 weeks or less, we are really interested in detecting the long-term effect of a change. How do long-term effects differ from short-term outcomes? How can we accurately measure those long- term factors without having to wait a long time in every case?" (Gupta et al. (2019), p. 21).

In these and many other examples, the researcher is faced with making recommendations regarding the future implementation of the intervention on the basis of measurements of its

effect on a variety of sometimes disparate and possibly conflicting outcome measures. A key question is how to balance these different outcomes when making an overall assessment. In practice, researchers often deemphasize short-term outcomes for which they do not find statistically significant effects, instead making perhaps somewhat *ad hoc* qualitative assessments regarding the relative importance of the remaining short-term outcomes.

In this paper we lay out a framework for analyzing these issues. We consider the scenario in which researchers do not measure the primary outcome in the context of data containing information on the intervention. Instead, we assume that the researcher has a second, observational, dataset where the researcher observes the surrogates and the primary outcome but does not observe the treatment. In both samples the researcher may also observe variables not affected by the treatment, such as pre-treatment characteristics of the participants.

We make four main contributions. First, we articulate three key assumptions under which the average effect of the treatment on the primary outcome is identified from the combination of the experimental and observational samples: (*i*) a standard assumption that the assignment in the experimental sample is *Unconfounded*; (*ii*) a *Surrogacy* assumption which requires that the causal path from the treatment to the primary outcome goes through the surrogates (Prentice, 1989; Day and Duffy, 1996; Begg and Leung, 2000; Frangakis and Rubin, 2002); and (*iii*) a *Comparability* or external validity assumption, which requires that the observational and experimental samples are comparable in the sense that the outcome distributions conditional on surrogates and pre-treatment variables are identical.

Under these three assumptions, the average effect of the treatment on the primary outcome can be estimated as the average effect of the treatment on an aggregate of the surrogates, which we label the surrogate index. This index combines the individual surrogates through their predicted value of the primary outcome. For example, when studying the impact of class size, the primary outcome might be high school graduation, while the two surrogates might be mathematics and reading scores. For the special case of linear models, the proposal boils down to multiplying the causal effects of the intervention on the two scores (which can be estimated

in the experimental sample) by the coefficients from a linear regression of the primary outcome on the two scores in the observational sample. The approach replaces a subjective assessment of the relative importance of the two short-term measures by an objective data-driven criterion, namely the predictive power of the scores for the outcome of interest.

In our second contribution, we derive the efficiency bound and propose various efficient estimators under various scenarios, including scenarios with a single sample or two samples, as well as with and without Surrogacy. This allows us to quantify the information content of the Surrogacy assumption.[1]

In our third contribution, we provide bounds on the biases that arise in scenarios where either or both of Surrogacy or Comparability are violated. We show that even if these assumptions fail to hold (but unconfoundedness does hold), the proposed estimators still estimate a well-defined causal effect, by providing a principled way of combining short-term outcomes in a single measure through their predicted effect on the long-term outcome.

In our fourth contribution, we evaluate these methods in the context of a labor market program where we observe long-term (thirty-six quarters) outcomes in four locations. Following an approach popularized by LaLonde (1986), we put aside part of the data and investigate whether we could have estimated the long-term effects without having long-term experimental data. Specifically, we take one of the locations, Riverside, and put aside the long-term outcome for individuals from that location. Then we take the other three locations, Alameda, Los Angeles, and San Diego, and put aside the treatment assignment for that sample. We investigate whether these two samples allow us to recover the experimental long-term effects in Riverside using surrogates corresponding to the first T quarters of outcomes (employment, earnings, and aid indicators). We find that combining six quarters of outcome data into a surrogate index suffices to obtain estimates close to the long run effects. Using the additional data that were put aside for the main analysis, we also directly test whether the critical assumptions, Surrogacy and Comparability, hold given various alternative sets of surrogates.

[1]We are grateful to Kevin Chen and David Ritzwoller for pointing out an error in one of our earlier efficiency bound calculations. See Chen and Ritzwoller (2023) for more details.

We recognize that the credibility of the Surrogacy assumption may be questioned in any given application, especially when viewed in isolation. Therefore, we view the best path forward as building a "library" of surrogate indices in which researchers systematically catalog across several studies the smallest set of surrogates that successfully match long-term outcomes of interest (*e.g.*, earnings, mortality, educational attainment). If one establishes, for instance, that six quarters of employment and earnings data are sufficient to predict the impacts of many different job training programs – as our cross-site comparisons of the GAIN program suggest – then the long-term impacts of future job training programs could be credibly estimated using the established six-quarter surrogate index. We view the empirical application in this paper as providing one element of such a library and hope future work will expand upon it by identifying surrogate indices that match estimated long-term impacts in other applications.

This study is related to three main bodies of literature, surrogacy, mediation, and missing data. We extend the literature on surrogacy (Prentice 1989; Day and Duffy 1996; Fleming and DeMets 1996; Begg and Leung 2000; Xu and Zeger 2001; Lauritzen 2004; D'Agostino, Campbell and Greenhouse 2006; Qu and Case 2006; Alonso et al. 2006; Gilbert and Hudgens 2008; Weir and Walley 2006) by formally including the presence of a second observational sample that is used to estimate the relationship between surrogates and the primary outcome and articulating the assumptions that justify doing so. In doing so we allow for uncertainty in the estimation of this surrogates/outcome relationship, whereas the previous literature took this relation as known. We also consider biases arising from violations of Surrogacy and Comparability.

In addition, this study builds on the literature on mediation (Baron and Kenny 1986; van der Laan and Petersen 2004; Imai, Keele and Tingley 2010; Zheng and van der Laan 2012; Tchetgen Tchetgen and Shpitser 2014; VanderWeele 2015), which considers the decomposition of an average treatment effect into the direct effect of a treatment on an outcome and indirect effects that flow through a mediator. In the mediation setup, all three key variables – the outcome, the treatment, and the mediator – are observed for the same units. The goal in the mediation literature is to determine the relative magnitudes of the direct and indirect effects. In our surrogacy

analysis we focus on the case in which the direct effect is absent by assumption.

This paper is also related to the classical missing data literature in statistics (Rubin 1976, 2004; Little and Rubin 2014). Our key assumptions are closely related to the Missing At Random (MAR) assumption. Our approach can be viewed as a special case of approaches that combine data sets, *e.g.*, Ridder and Moffitt (2007); Chen et al. (2008). In particular Rässler (2004, 2012) refers to our setting, where one variable is missing in one part of a sample and a second variable missing in the remainder of the sample, as a "data fusion" setting. Graham, Pinto and Egel (2016) discuss efficient estimation for a particular set of models defined by moment conditions in such a data fusion setting, where they allow the treatment to be a general random variable, rather than a binary indicator as in our setup.

The paper is organized as follows. Section 2 sets up the problem and introduces the notation. Section 3 discusses the critical assumptions and links the setup to the mediation and missing data literature. Section 4 discusses identification and the efficiency bounds. Section 5 presents formulas for bias when the surrogacy assumption fails and derives bounds on the degree of bias. Section 6 discusses estimation. Section 7 presents the empirical application. Section 8 concludes.

2 Setup and Notation

We define two samples, an Experimental (E) sample and an Observational (O) sample, with $N_{\rm E}$ and $N_{\rm O}$ units or individuals, respectively. It is convenient to view the data as consisting of a single sample of size $N = N_{\rm E} + N_{\rm O}$, with $P_i \in \{\rm O, E\}$ a binary indicator denoting the sample to which unit i belongs.

For each unit, there is a binary treatment of interest, $W_i \in \{0, 1\}$, and a scalar primary outcome, denoted by Y_i. This outcome is not observed for individuals in the experimental sample. In addition, there are intermediate or secondary outcomes, which we refer to as surrogates (to be defined precisely in Section 3.2), denoted by S_i for each unit. Typically, the surrogate outcomes are vector-valued in order to make the properties we define plausible. Finally, we measure pre-treatment covariates X_i for each unit, known not to be affected by the treatment.

Following the potential outcomes framework or Rubin Causal Model (Rubin 1974; Holland 1986; Imbens and Rubin 2015), individuals in this group have two pairs of potential outcomes: $(Y_i(0), Y_i(1))$ and $(S_i(0), S_i(1))$. The realized outcomes are related to their respective potential outcomes as follows.

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases} \quad \text{and} \quad S_i \equiv S_i(W_i) = \begin{cases} S_i(0) & \text{if } W_i = 0, \\ S_i(1) & \text{if } W_i = 1. \end{cases}$$

Overall, the units are characterized by the values of the septuple $(Y_i(0), Y_i(1), S_i(0), S_i(1), X_i, W_i, P_i)$. We do not observe the full septuple for any units. Rather, for units in the experimental sample we observe the triple (X_i, W_i, S_i) with support $(\mathbb{X}, \mathbb{W}, \mathbb{S})$ where $\mathbb{W} = \{0, 1\}$. In the observational sample, we do not observe to which treatment each of the $N_{\rm O}$ individuals were assigned. We observe the triple (X_i, S_i, Y_i), with support $\mathbb{X}$, $\mathbb{S}$, and $\mathbb{Y}$ respectively. To simplify the exposition, we analyze the data as if we have a random sample from a population of units for which we observe the quintuple $(P_i, X_i, S_i, \mathbf{1}_{P_i={\rm E}} W_i, \mathbf{1}_{P_i={\rm O}} Y_i)$, where we treat P_i as a random variable taking on the values $\{{\rm O}, {\rm E}\}$.

Assumption 1. *We have a single random sample of size N drawn from the joint distribution of $(P_i, X_i, S_i, W_i, Y_i)$, where we observe for each unit in the sample $(P_i, X_i, S_i, \mathbf{1}_{P_i={\rm E}} W_i, \mathbf{1}_{P_i={\rm O}} Y_i)$.*

We summarize this data setup in Table 1. The setup differs from those in Athey, Chetty and Imbens (2020) and Kallus and Mao (2020), where we would also observe the treatment in the observational sample, but in the experimental sample we would still not observe the primary outcome.

We are interested in the Average Treatment Effect (ATE) on the primary outcome in the population from which the experimental sample is drawn:

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) | P_i = {\rm E}]. \tag{2.1}$$

The same issues we study in the current paper apply to other estimands, such as the average treatment effect for the treated units, or the average for the observational sample.

Table 1: Observation Scheme: ✓ is observed, ? is missing

Units	Sample P_i	Treatment W_i	Long-Term Outcome Y_i	Surrogate S_i	Pretreatment Variables X_i
1 to $N_{\rm E}$	E	✓	?	✓	✓
$N_{\rm E}+1$ to $N_{\rm E}+N_{\rm O}$	O	?	✓	✓	✓

An implicit assumption in our setup is that the two variables that are common to both samples, S_i and X_i, measure the same underlying variables in both samples. In some cases it is possible that in one of the two samples, a coarser version is measured, for example age or education may be measured in multi-year categories rather than in years. In that case, a simple solution is to proceed by using the coarser version of the variables as corresponding to the surrogate of pre-treatment variable. Another complication arises if the unit of observation differs in two samples, say individuals versus zipcodes. Again additional assumptions are required to link the variables between samples.

Table 2 summarizes key definitions and notation.

3 The Critical Assumptions: Unconfoundedness, Surrogacy, and Comparability

In this section, we discuss the three key assumptions that together allow us to combine the observational and experimental samples and estimate the causal effect of the treatment on the primary outcome, exploiting the presence of the surrogates. The first assumption is *Unconfoundedness* or *Ignorability*, common in the program evaluation literature (Rosenbaum and Rubin 1983*b*; Imbens and Rubin 2015), which ensures that adjusting for pre-treatment variables leads to valid causal effects in the experimental sample. The second assumption is the *Surrogacy condition*

Table 2: Notation and Definitions

Sampling Indicator	$P_i \in \{\text{E}, \text{O}\}$
Potential Outcomes for Primary Outcome	$Y_i(0), Y_i(1)$
Potential Outcomes for Surrogates	$S_i(0), S_i(1)$
Binary Treatment Indicator	$W_i \in \{0, 1\}$
Realized Value for Outcome	$Y_i = Y_i(W_i)$
Realized Value for Surrogate	$S_i = S_i(W_i)$
Estimand	$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid P_i = \text{E}]$
Surrogate Index	$\mu(s, w, x, p) \equiv \mathbb{E}[Y_i \mid S_i = s, W_i = w, X_i = x, P_i = p]$ $\mu(s, x, p) \equiv \mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = p]$ $\mu(w, x) \equiv \mathbb{E}[\mu(S_i, X_i, O) \mid W_i = w, X_i = x, P_i = E]$ $\sigma^2(s, w, x, p) \equiv \mathbb{V}(Y_i \mid S_i = s, W_i = w, X_i = x, P_i = p)$ $\sigma^2(s, x, p) \equiv \mathbb{V}(Y_i \mid S_i = s, X_i = x, P_i = p)$ $\sigma^2(w, x) \equiv \mathbb{V}(Y_i \mid W_i = w, X_i = x, P_i = \text{O})$
Surrogate Score	$\rho(s, x) \equiv \text{pr}(W_i = 1 \mid S_i = s, X_i = x, P_i = \text{E})$
Propensity Score	$\rho(x) \equiv \text{pr}(W_i = 1 \mid X_i = x, P_i = \text{E})$ $\rho \equiv \text{pr}(W_i = 1 \mid P_i = \text{E})$
Sampling Score	$\varphi(s, x) \equiv \text{pr}(P_i = \text{E} \mid S_i = s, X_i = x)$ $\varphi(x) \equiv \text{pr}(P_i = \text{E} \mid X_i = x)$ $\varphi \equiv \text{pr}(P_i = \text{E})$
Conditional Distribution of Surrogates	$\pi(s \mid w, x) \equiv f_{S_i \mid W_i, X_i, P_i}(s \mid w, x, \text{E})$ $\pi(s \mid x) \equiv f_{S_i \mid X_i, P_i}(s \mid x, \text{E})$
Influence Function	$\psi(y, s, w, x, p)$

Notes: This table summarizes the notation. Conditional expectations and variances of the outcome Y_i are denoted by $\mu(\cdot)$ and $\sigma^2(\cdot)$ respectively. Conditional probabilities of the treatment are denoted by $\rho(\cdot)$. Conditional probabilities of the sample are denoted by $\varphi(\cdot)$ The arguments of these functions can be both the surrogates S_i and the pre-treatment variables X_i, or just the pre-treatment variables X_i.

due to Prentice (1989), that allows us to use the surrogate variables to proxy for the primary outcome. The third assumption is *Comparability*, which formalizes the connection between the two samples. This assumption is rarely stated formally, but plays an important role in our analysis.

3.1 Unconfoundedness

For the individuals in the experimental group, the propensity score is the conditional probability of receiving the treatment: $\rho(x) \equiv \text{pr}(W_i = 1 | X_i = x, P_i = \text{E})$. We assume that for individuals in the experimental group, treatment assignment is unconfounded, and we have overlap in the distribution of pre-treatment variables between the treatment and control groups (Rosenbaum and Rubin 1983*b*; Imbens and Rubin 2015):

Assumption 2. (Unconfounded Treatment Assignment / Strong Ignorability)
(*i*)

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1), S_i(0), S_i(1)\Big) \;\Big|\; X_i, P_i = \text{E},$$

(*ii*) $0 < \rho(x) < 1$ for all $x \in \mathbb{X}$.

This assumption, widely used in the causal inference literature, implies that in the experimental sample, we can estimate the average causal effect of the treatment on the surrogates by adjusting for pre-treatment variables. We would also have been able to estimate the causal effect on the primary outcome had the primary outcome been measured in the experimental sample. In many applications of surrogacy approaches, the treatment in the experimental sample is assigned completely randomly. In that case this assumption is satisfied by design. However, unconfoundedness is all that is required.

3.2 Surrogacy

Next we discuss the second critical assumption, surrogacy. We also introduce two concepts, the *surrogacy score*, similar to the propensity score, and the *surrogacy index*, to combine multiple surrogates.

3.2.1 The Prentice Criterion

Prentice 1989 defines a surrogate as a post-treatment variable where conditioning on it makes the outcome and the treatment independent:

Assumption 3. (Surrogacy, Prentice Criterion)
(i)

$$W_i \perp\!\!\!\perp Y_i \,\Big|\, S_i, X_i, P_i = \mathrm{E}.$$

and (ii) $0 < \rho(s,x) < 1, \text{for all } s \in \mathbb{S}, x \in \mathbb{X},$ *and* $0 < \mathrm{pr}(P_i = \mathrm{E}) < 1.$

Remark 1. *If the quadruple* (Y_i, S_i, W_i, X_i) *were observed for all units, surrogacy would be a testable condition. With* (S_i, W_i, X_i) *observed for units in the experimental sample, and* (Y_i, S_i, X_i) *observed for units in the observational sample, this assumption has no testable implications.*

Remark 2. *Note that Surrogacy is formulated in terms of the realized outcome and surrogate values. In contrast we formulated the ignorability condition (Assumption 2) in terms of the potential outcomes. This is partly to connect our discussion to the surrogacy literature (Prentice, 1989; Day and Duffy, 1996).*

Surrogacy is often debated in empirical applications. Freedman, Graubard and Schatzkin (1992) argue that the surrogate may not mediate the full effect of the treatment in many settings. For example, reductions in class size may affect earnings through changes in non-cognitive skills that are not fully captured by standardized test scores (Heckman, Stixrud and Urzua 2006; Chetty et al. 2011).

3.2.2 The Surrogacy Index and the Surrogacy Score

There are two scalar functions of the surrogates that play an important role in the analyses: the surrogate index and surrogate score.

Definition 1. (The Surrogate Index) *The surrogate index is the conditional expectation of the primary outcome given the surrogate outcomes and the pre-treatment variables, conditional on the sample:*

$$\mu(s,x,p) \equiv \mathbb{E}\left[Y_i \middle| S_i = s, X_i = x, P_i = p\right].$$

Remark 3. *The surrogate index in the observational sample, $\mu(s,x,\mathrm{O})$, is identified because we observe the triple (Y_i, S_i, X_i) in the observational sample.*

Definition 2. (The Surrogate Score) *The surrogate score is the conditional probability of having received the treatment given the value for the surrogate outcomes and the covariates in the experimental sample:*

$$\rho(s,x) \equiv \mathrm{pr}(W_i = 1 | S_i = s, X_i = x, P_i = E).$$

The surrogacy score plays is similar to the role the propensity score plays in analyses under unconfoundedness (Rosenbaum and Rubin, 1983*b*). Here if the surrogacy condition holds conditional on (S_i, X_i), it also holds conditional on the surrogacy score.

Proposition 1. (Surrogate Score) *Suppose Surrogacy (Assumption 3) holds. Then:*

$$W_i \perp\!\!\!\perp Y_i \,\Big|\, \rho(S_i, X_i), P_i = \mathrm{E}.$$

All proofs are given in the Appendix.

3.2.3 The Benefits of Multiple Surrogates

One theme of this paper is that having multiple short-term variables can make a surrogacy approach more plausible, the same way multiple pre-treatment variables can make the unconfoundedness assumption more plausible. Here we discuss some illustrative examples.

The first example is illustrated in Figure 1.A. Suppose the treatment is an educational intervention. This treatment affects the outcome of interest, some labor market outcome, e.g., earnings, through a number of different channels corresponding to different skill sets. These

channels may include mathematics skills, language skills, and social skills. Using only one of these variables as a surrogate would lead to biased estimates because they would ignore the other causal paths. In this case the set of three short-term variables collectively satisfy Unconfoundedness and Surrogacy.

The second case is illustrated in Figure 1.B. In this setup there is a variable, labeled "skills', that satisfies the critical assumptions for surrogacy. However, skills is not observed by the researcher. Instead we have two noisy measures of this surrogate, say both a written and an oral exam. Collectively these two variables may still not satisfy Surrogacy, since there may be impacts of skills on earnings not captured by the exams, but the bias from using both would be less than the bias from using only one candidate surrogate.

Figure 1.a Surrogacy Assumption Satisfied

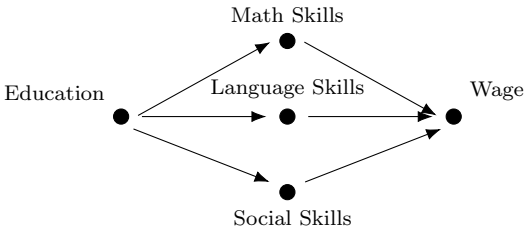


Figure 1.c: Multiple Surrogates, Scenario 1

Figure 1.b Multiple Surrogates

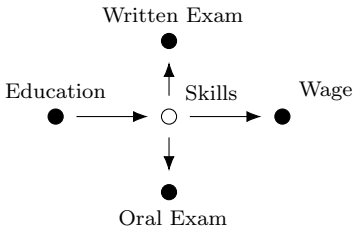
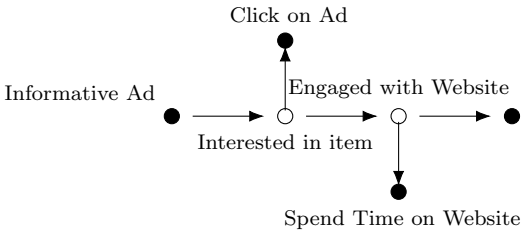
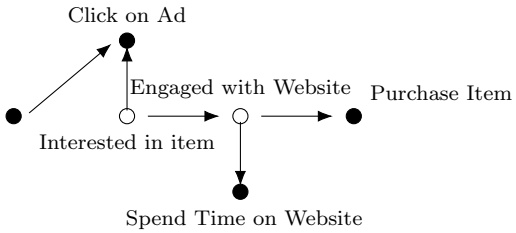


Figure 1.d: Multiple Surrogates, Scenario 2





The third case is illustrated in Figure 1.C. Here there is a pathway from the treatment, an informative advertisement about an item, to the outcome, an indicator for the individual purchasing the advertised item, going through two variables that on their own could each serve as surrogates. These two variables are whether someone has interested in the item, and whether the individual engaged with the website where the item was sold. However, we only measure

noisy versions of these surrogates. For the first surrogate we observe whether an individual clicked on the advertisement for the item, and for the second surrogate we observe the time spent on the website. Neither of these two observed variables is a valid surrogate, but the combination of the two generally removes more of the bias than a single one.

Figure 1.D illustrates further the concerns with only using a single surrogate, the possibility of focusing on treatments that improve the surrogate variable but not the primary outcome. Suppose that a researcher uses the single variable "click on ad" as a surrogate for the effect of the ad on purchases. If the observational sample was based on informative advertisements, there is likely a positive correlation between clicking on the advertisement and purchases. However, if the new treatment is uninformative, *e.g.,* clickbait advertisement, with no effect on the actual interest in the item, the surrogacy analysis using click behavior as the surrogate will be ineffective. Using both click behavior and time spent on the website as surrogates will likely reduce the bias. The same argument implies that using multiple tests as surrogates can reduce problems with "teaching to the test," where the long-run impact of an intervention is not well captured by scores on a test.

3.3 Comparability

Surrogacy and Unconfoundedness by themselves are not sufficient for consistent estimation of τ because they do not place restrictions on how the relationship between Y_i and S_i in the observational sample compares to that in the experimental sample. As far as we know, such restrictions were not previously articulated in the surrogacy literature because the setup is typically one with just the separate experimental sample. However, a comparability assumption is implicit in the way the postulated relationship between the surrogate and the primary outcome is used in that literature. Related assumptions about the possibility of using causal estimates in one location to predict causal effects in a second location on the basis of distributions of pre-treatment variables are discussed in Hotz, Imbens and Mortimer (2005) and the literature on transportability, Pearl and Bareinboim (2014).

3.3.1 The Comparability Assumption

Let $\varphi \equiv \text{pr}(P_i = \text{E})$ be the probability of a unit being part of the experimental sample. We introduce the *Sampling Score,* the propensity to be in the experimental sample:

Definition 3. (Sampling Score)
The sampling score is $\varphi(s,x) \equiv \text{pr}(P_i = \text{E}|S_i = s, X_i = x)$.

The third key assumption we make is that the conditional distribution of Y_i given (S_i, X_i) in the observational sample is the same as the conditional distribution of Y_i given (S_i, X_i) in the experimental sample, and that the support of (S_i, X_i) in the experimental sample is a subset of that in the observational sample. Formally,

Assumption 4. (Comparability of Samples)
(i) $P_i \perp\!\!\!\perp Y_i \,\Big|\, S_i, X_i,$
(ii) $\varphi(s,x) < 1$ for all $s \in \mathbb{S}$ and $x \in \mathbb{X}$.

Similar to Unconfoundedness and Surrogacy this is a strong assumption, but unlike those assumptions it is rarely discussed explicitly. As we show in Section 5, by making it explicit we can discuss the biases arising from violations and improve the intuition when this assumption may be of concern. If the observational and experimental samples are substantially different in terms of the distribution of pre-treatment variables and surrogates, it would likely be more controversial to assume that conditional on those variables the outcome distributions are identical.

3.3.2 The Surrogate Index and the Sampling Score

We let $\mu(s,w,x,p)$ denote the conditional expectation of the primary outcome given pre-treatment variables, surrogates, treatment, and sample:

$$\mu(s,w,x,p) \equiv \mathbb{E}\left[Y_i \middle| S_i = s, X_i = x, W_i = w, P_i = p\right]. \tag{3.1}$$

Comparability and Surrogacy together allow us to impute the missing primary outcomes in the experimental sample, as shown by the following proposition.

Proposition 2. (SURROGATE INDEX) (*i*) *Suppose Assumption 3 (Surrogacy) holds. Then:*

$$\mu(s,w,x,\mathrm{E}) = \mu(s,x,\mathrm{E}), \qquad \text{for all } s \in \mathbb{S},\ x \in \mathbb{X}, \quad \text{and } w \in \mathbb{W}.$$

(*ii*) *Suppose Assumption 4 (Comparability) holds. Then:*

$$\mu(s,x,\mathrm{E}) = \mu(s,x,\mathrm{O}) \quad \text{for all } s \in \mathbb{S}, \text{ and } x \in \mathbb{X}.$$

(*iii*) *Suppose Assumptions 3 (Surrogacy) and 4 (Comparability) hold. Then:*

$$\mu(s,w,x,\mathrm{E}) = \mu(s,x,\mathrm{O}) \quad \text{for all } s \in \mathbb{S}, x \in \mathbb{X}, \text{ and } w \in \mathbb{W}.$$

Because we can estimate $\mu(s,x,\mathrm{O}) = \mathbb{E}[Y_i|S_i = s, X_i = x, P_i = \mathrm{O}]$, we can impute the missing Y_i in the experimental sample as $\mu(S_i, X_i, \mathrm{O})$.

3.4 Surrogacy, Mediation, Instrumental Variables, Directed Acyclical Graphs, and Missing Data

To provide context for the setup here and the key assumptions, it is useful to make a link to three related literatures, on mediation, instrumental variables, and missing data respectively. We describe the causal structures for surrogacy, mediation, and instrumental variables using a directed acyclical graph (DAG) (Pearl, 2000). The interpretations provided in this subsection are note essential to the main results in the next section.

3.4.1 Directed Acyclical Graph Representations

The surrogacy, mediation, and instrumental variables literatures all study causal structures involving a causally linked sequence of three (sets) of variables. They differ in three key aspects: (*i*) the assumptions they make on the causal structure, (*ii*) the estimands that are the primary focus of the analysis, and (*iii*) the data available for the analyses. The literatures also differ in the labels typically used for the three variables. In Table 3 we list the labels, estimands, and some of the assumptions.

Table 3: Surrogacy versus Mediation versus Instrumental Variables

	Surrogacy	Mediation	Instrumental Var
Left Variable (L)	Treatment (W)	Treatment (W)	Instrument (Z)
Middle Variable (M)	Surrogate (S)	Mediator (M)	Treatment (W)
Right Variable (R)	Outcome (Y)	Outcome (Y)	Outcome (Y)
Estimand	Effect of L on R	Direct and Indirect Effect of L on R	Effect of M on R
Direct Effect of L on R	No	Yes	No
Unobs Conf between L and M	No	No	No
Unobs Conf between M and R	No	No	Yes
All Variables Observed Together	No	Yes	Yes

In Figures 2.A-2.C we show the differences in structures in DAG form in a single sample setting (so that we need not be concerned with the comparability assumption). Figure 2.A illustrates the surrogacy setup, with a causal link from the treatment to the surrogate and from the surrogate to the outcome. There is no unobserved confounder for the causal relation between treatment and surrogate, which would violate Assumption 2 (Unconfounded Treatment Assignment / Strong Ignorability). There is no direct causal link from the treatment to the outcome. There are also no unobserved confounders for the causal relation between surrogate and the outcome. These two features of the DAG (no direct link between treatment and outcome and no unobserved confounder for the relation between surrogate and outcome imply Assumption (3) (Surrogacy).

Figure 2.B shows a mediation example where Assumption 3 is violated because there is a direct effect of the treatment on the outcome that does not pass through the surrogate. In this case S_i is a typically labelled a mediator, rather than a surrogate. In the mediation case the direct effect of the treatment on the outcome is estimable because all three variables, treatment, mediator and outcome are observed in the same sample.

Figure 2.A. Surrogacy Assumption Satisfied

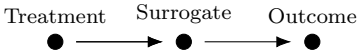


Figure 2.B. Violation of Surrogacy due to Direct Effect (Mediation Setup)

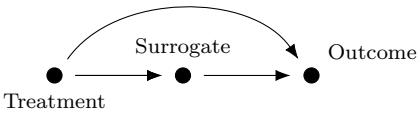


Figure 2.C. Violation of Surrogacy Assumption due to Unobserved Confounder (IV Setup)

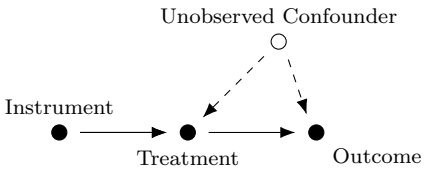


Figure 2.C shows a DAG representation of the standard instrumental variables (IV) model familiar to economists. The first difference from the surrogacy setup in Figure 2.A is that in the instrumental variables setting the interest is in the causal effect of the variable in the middle of the three variable chain (the surrogate S in the surrogacy setting, and the treatment W in the instrumental variables setting), on the outcome, whereas in the surrogacy setting the primary interest is in the effect of the first variable in the chain (the treatment W in the surrogacy setting and the instrument Z in the instrumental variables setting) on the outcome. In the instrumental variables case the surrogacy estimand is immediately identified as the intention-to-treat effect of the instrument, since the instrument and the surrogate are observed in the same sample. Under the assumptions of the surrogacy setup, the target for an instrumental variables analysis, the effect of the surrogate on the primary outcome, is immediately identified. The instrumental variables settings is characterized by the presence of an unobserved confounder that affects both the treatment of interest and the outcome. The presence of that unobserved confounder violates Surrogacy, even if the treatment has no direct effect on the long-term outcome (Frangakis and Rubin 2002; Rosenbaum 1984; Joffe and Greene 2009; VanderWeele 2015).

The presence of this unobserved confounder also violates the comparability assumption if the marginal distribution of the treatment W differs between the observational and experimental samples, as will typically be the case. In both the surrogacy and the instrumental variables cases, we assume the absence of a direct effect of the first variable in the causal chain (the treatment W in the surrogacy case and the instrument in the instrumental variables case) on the primary outcome. In the surrogacy setting, this assumption is part of the Surrogacy assumption, while

in the instrumental variables setting this is typically referred to as the exclusion restriction (Angrist, Imbens and Rubin, 1996).

3.4.2 A Missing Data Representation

In the Online Appendix we also discuss a missing data interpretation of the surrogacy approach. Eessentially we show that the following joint conditional independence assumption,

$$P_i \perp\!\!\!\perp Y_i \perp\!\!\!\perp W_i \,\Big|\, S_i, X_i, \tag{3.2}$$

implies both surrogacy and comparability.

This missing data characterization is useful because it allows one to use insights from the missing data literature, both for the current problems and for generalizations. Given (3.2) we can use the conditional distribution of Y_i given (S_i, X_i) in the observational sample with $P_i = \text{O}$ to impute the missing outcomes in the experimental sample with $P_i = \text{E}$, and we can use the conditional distribution of W_i given (S_i, X_i) in the experimental sample wth $P_i = \text{E}$ to impute the missing treatments in the observational sample with $P_i = \text{O}$.

This observation directly extends to more general imputation problems. Suppose we have two samples where in one sample, indicated by $P_i = \text{E}$ we observe one set of variables, (Z_{i1}, Z_{i2}) and in the second sample, indicated by $P_i = \text{O}$ we observe a partially overlapping set of variables, (Z_{i2}, Z_{i3}). Then the analogous assumption that allows the imputation of all missing variables is $P_i \perp\!\!\!\perp Z_{i1} \perp\!\!\!\perp Z_{i3} | Z_{i2}$.

4 Identification and Semiparametric Efficiency Bounds

4.1 Three Identification Results

We now present our central identification result. We analyze three different representations of the average treatment effect that lead to three estimation strategies, somewhat similar to inverse propensity score weighting, regression, and influence function estimators for average treatment effects under unconfoundedness (Imbens, 2004). The motivation for developing the

different representations is that estimators corresponding to those different representations can have different properties in finite samples, just like they do in the unconfoundedness setting. Estimators based on the first representation require estimation of the surrogate index, but not the surrogate score. Estimators based on the second representation instead require estimation of the surrogate score, but not the surrogate index. Estimators based on the third representation require estimation of both, but have attractive double robustness properties.

We define the following four objects, all functionals of distributions that are directly estimable from the data. First define the statistical estimand, the average difference in the surrogate index between treated and control, adjusted for pretreatment variablles, in the experimental sample:

$$\tau^* \equiv \mathbb{E}\left[\left\{\mathbb{E}\Big[\mathbb{E}\left[Y_i|\, S_i, X_i, P_i = \mathrm{O}\right]\Big| W_i = 1, X_i, P_i = \mathrm{E}\Big]\right. \tag{4.1}$$

$$\left. -\mathbb{E}\Big[\mathbb{E}\left[Y_i|\, S_i, X_i, P_i = \mathrm{O}\right]\Big| W_i = 0, X_i, P_i = \mathrm{E}\Big]\right\}\Big|\, P_i = \mathrm{E}\Big].$$

Next, with a surrogate index representation:

$$\tau^{\mathrm{E}} \equiv \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1 - W_i}{1 - \rho(X_i)}\,\middle|\, P_i = \mathrm{E}\right], \tag{4.2}$$

then a surrogate score representation,

$$\tau^{\mathrm{O}} \equiv \mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1 - \varphi)}{\rho(X_i) \cdot (1 - \varphi(S_i, X_i)) \cdot \varphi}\right. \tag{4.3}$$

$$\left. - Y_i \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \varphi(S_i, X_i) \cdot (1 - \varphi)}{(1 - \rho(X_i)) \cdot (1 - \varphi(S_i, X_i)) \cdot \varphi}\,\middle|\, P_i = \mathrm{O}\right].$$

The third representation is based on the influence function. We first define

$$\mu(w, x) \equiv \mathbb{E}[\mu(S_i, X_i, \mathrm{O})|W_i = w, X_i = x, P_i = \mathrm{E}].$$

Then the influence function is

$$\psi(y, s, w, x, p) = \frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\left(\frac{w \cdot (\mu(s, x, \mathrm{O}) - \mu(1, x))}{\rho(x)} - \frac{(1 - w) \cdot (\mu(s, x, \mathrm{O}) - \mu(0, x))}{1 - \rho(x)}\right) \tag{4.4}$$

$$+\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\Big(\mu(1, x) - \mu(0, x) - \tau\Big)$$

$$+\frac{\mathbf{1}_{p=\mathrm{O}}}{\varphi}\frac{\varphi(s,x)}{1-\varphi(s,x)}\frac{\left(y-\mu(s,x,\mathrm{O})\right)\left(\rho(s,x)-\rho(x)\right)}{\rho(x)(1-\rho(x))}$$

with the estimand

$$\tau^{\mathrm{O,E}}=\mathbb{E}\left[\psi(Y_i,S_i,W_i,X_i,P_i)+\tau\right]. \tag{4.5}$$

Remark 4. *An earlier version of the paper had a mistake in the representation of the influence function. We are grateful to Kevin Chen and David Ritzwoller for pointing this out. See Chen and Ritzwoller (2023) for details.*

Theorem 1. (IDENTIFICATION) *(i) Suppose that Assumption 1 holds. Then, assuming all expectations are finite,*

$$\tau^*=\tau^{\mathrm{E}}=\tau^{\mathrm{O}}=\tau^{\mathrm{O,E}}.$$

(ii) Suppose that Assumptions 1–4 hold. Then the average treatment effect is equal to the following three estimable functions of the data:

$$\tau\equiv\mathbb{E}[Y_i(1)-Y_i(0)|P_i=\mathrm{E}]=\tau^*=\tau^{\mathrm{E}}=\tau^{\mathrm{O}}=\tau^{\mathrm{O,E}},$$

(iii) Jointly Assumptions 1, 2(i), 3(i) and 4(i) have no testable implications.

Remark 5. *The first part of the theorem implies that the four functionals of the joint distribution of $(\mathbf{1}_{P_i=\mathrm{O}}Y_i,S_i,\mathbf{1}_{P_i=\mathrm{E}}W_i,X_i,P_i)$ are identical, irrespective of the Unconfoundedness, Surrogacy, and Comparability assumptions.*

Remark 6. *Just like in the unconfoundedness case (Newey, 1994; Chernozhukov et al., 2016), the influence function representation is doubly robust. Chen and Ritzwoller (2023) show that if the functions in the influence function that represent conditional expectations of the outcome, $\mu(s,x,\mathrm{O})$ and $\mu(w,x)$, are correctly specified, then the influence function has expectation zero irrespective of the functions used for the various propensity score, $\rho(s,x)$, $\rho(x)$, ρ, and the sampling score $\varphi(s,x)$. Similarly, if the various propensity score, $\rho(s,x)$, $\rho(x)$, ρ, and the sampling score $\varphi(s,x)$ are correct, the influence function has expectation zero, irrespective of the functions used for the conditional outcome expectations $\mu(s,x,\mathrm{O})$ and $\mu(w,x)$.*

4.2 Semiparametric Efficiency Bounds

In this subsection we present two pairs of semiparametric efficiency bound results (Bickel et al., 1993; Newey, 1990) for two different data configurations. The first directly refers to the main setup in this paper with the experimental and observational sample. This result is essentially shown in Chen and Ritzwoller (2023) which corrects a mistake in an earlier version of the current paper.

Theorem 2. *Suppose Assumptions 1–4 hold. Then*
(i) the semiparametric efficiency bound, normalized by the square root of the sample size N, is

$$\begin{aligned}\mathbb{V} &= \mathbb{E}[\psi(Y_i, S_i, W_i, X_i, P_i)^2] \\ &= \mathbb{E}\Bigg[\frac{1-\varphi(S_i,X_i)}{\varphi^2}\left(\left(\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})\right) \\ &+\frac{\varphi(X_i)}{\varphi^2}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2 \\ &+\frac{\varphi(S_i,X_i)}{\varphi^2}\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg].\end{aligned}$$

(ii) If in addition the observational sample is large relative to the experimental sample, and $\sup_{s,x}\varphi(s,x)\to 0$, then the efficiency bound, now normalized by the expected sample size of the experimental sample, $\mathbb{E}[N_{\mathrm{E}}]=\varphi N$ simplifies to

$$\mathbb{E}\left[(\mu(1,X_i)-\mu(0,X_i)-\tau)^2+\frac{(1-W_i)(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{W_i(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\,\middle|\,P_i=\mathrm{E}\right].$$

Remark 7. *This variance in part (ii) of Theorem 2 is smaller than the effiency bound we would obtain in a randomized experiment where we do observe the primary outcome and did not observe the surrogate. The bound in that case is well known since Hahn (1998),*

$$\mathbb{E}\left[(\mu(1,X_i)-\mu(0,X_i)-\tau)^2+\frac{(1-W_i)(Y_i-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{W_i(Y_i-\mu(1,X_i))^2}{\rho(X_i)^2}\,\middle|\,P_i=\mathrm{E}\right].$$

This advantage in terms of asymptotic precision of using the (true) predicted outcome $\mu(S_i, X_i, \mathrm{O})$ rather than the actual outcome Y_i has been noted previously in Day and Duffy (1996) in a setting with binary outcomes. In the general case this gain is equal to

$$\mathbb{E}\left[\frac{(1-W_i)(Y_i-\mu(S_i,X_i,\mathrm{O}))^2}{(1-\rho(X_i))^2}+\frac{W_i(Y_i-\mu(S_i,X_i,\mathrm{O}))^2}{\rho(X_i)^2}\middle| P_i=\mathrm{E}\right].$$

Next we consider the case where in a single sample we observe the treatment, primary outcome, surrogates and pre-treatment variables. In this single sample case we do not need the fifth variable, $P_i \in \{\mathrm{E}, \mathrm{O}\}$. To maintain consistency with the other parts of the discussion and to avoid ambiguity, we keep the notation as before. In this case we can think of P_i always taking the value $P_i = \mathrm{E}$. We calculate the efficiency bound both without the assumption that surrogacy holds and with the assumption that surrogacy holds. We do so for a data generating process where surrogacy does hold, to see the information gain from that assumption.

Theorem 3. *Suppose Assumptions 2 and 3 hold. (i) The variance bound without assuming surrogacy is*

$$\begin{aligned}\mathbb{V}_{\mathrm{ns}} = \mathbb{E}\Bigg[&\sigma^2(S_i,X_i,\mathrm{E})\cdot\left(\frac{\rho(S_i,X_i)}{\rho(X_i)^2}+\frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\right)+(\mu(1,X_i)-\mu(0,X_i)-\tau)^2\\&+\frac{\rho(S_i,X_i)}{\rho(X_i)^2}\cdot(\mu(S_i,X_i,\mathrm{E})-\mu(1,X_i))^2+\frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\cdot(\mu(S_i,X_i,\mathrm{E})-\mu(0,X_i))^2\Bigg].\end{aligned}$$

(ii) The efficiency gain from assuming surrogacy is

$$\Delta = \mathbb{V}_{\mathrm{ns}} - \mathbb{V}_{\mathrm{s}} = \mathbb{E}\left[\sigma^2\left(S_i, X_i, E\right)\frac{\rho\left(S_i,X_i\right)\left(1-\rho\left(S_i,X_i\right)\right)}{\rho\left(X_i\right)^2\left(1-\rho\left(X_i\right)\right)^2}\right] \geq 0,$$

where $\mathbb{V}_{\mathrm{s}}$ is the variance bound for the case with surrogacy,

$$\begin{aligned}\mathbb{V}_{\mathrm{s}} = \mathbb{E}\Bigg[&\sigma^2(S_i,X_i,\mathrm{E})\left(\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2+(\mu(1,X_i)-\mu(0,X_i)-\tau)^2\\&+\frac{\rho(S_i,X_i)}{\rho(X_i)^2}(\mu(S_i,X_i,\mathrm{E})-\mu(1,X_i))^2+\frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}(\mu(S_i,X_i,\mathrm{E})-\mu(0,X_i))^2\Bigg]\end{aligned}$$

Remark 8. *The expression for $\mathbb{V}_{\mathrm{ns}}$ is equivalent to the efficiency bound in (Hahn, 1998). It is written here in terms of the surrogates to facilitate the comparison to the efficiency bound exploiting surrogacy.*

Remark 9. *Note that the variance bound in part* (ii) *of Theorem 3 differs from that in Theorem 2*(ii) *which was derived under the same surrogacy assumption, but assuming that the observational sample was infinitely large, so the relation between the surrogates and the primary outcome was known without error. The result in* (ii) *captures just the value of the surrogacy assumption.*

5 Violations of the Surrogacy and Comparability Assumptions: Biases and Bounds

The three critical assumptions, Unconfoundedness, Surrogacy, and Comparability, are strong. There is a large literature studying the sensitivity to unconfoundedness conditions (Rosenbaum and Rubin, 1983*a*; Imbens, 2003; Cinelli and Hazlett, 2020) or bounds (Manski, 1990). Multiple studies have also raised concerns that in practice Surrogacy may not be satisfied (Begg and Leung, 2000; Freedman, Graubard and Schatzkin, 1992; Frangakis and Rubin, 2002; Rosenbaum, 1984; Joffe and Greene, 2009; VanderWeele, 2015), although we are not aware of formal sensitivity or bounds analyses. Violations of Comparability have not been explored because this assumption has not been previously formalized. In this section we examine the biases that arise from violations of Surrogacy and Comparability. We first characterize these biases and then derive estimable bounds on the magnitude of the biases that can arise from such violations.

5.1 Biases

We begin by characterizing the probability limit of estimators based on the representations of the estimand, τ^{E}, τ^{O}, and $\tau^{\mathrm{O,E}}$, in Theorem 1 when the Surrogacy and Comparability assumptions are violated, as well as in cases where the surrogate index is misspecified. Throughout the section, we maintain Unconfoundedness in the experimental sample (Assumption 2), and the random sampling assumption (Assumption 1). We denote the probability limit of the estimators by $\underline{\tau}$ to differentiate it from the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)|P_i = \mathrm{E}]$.

Theorem 4. (i) *Suppose Assumption 2 (Unconfoundedness) holds, but Assumptions 3 (Surro-*

gacy) and 4 (Comparability) do not necessarily hold. Then

$$\underline{\tau} \equiv \tau^{\mathrm{O}} = \tau^{\mathrm{E}} = \tau^{\mathrm{E,O}} = \mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) - \mu(S_i(0), X_i, \mathrm{O}) \middle| P_i = \mathrm{E}\right].$$

(ii) Suppose Assumptions 2 (Unconfoundedness) and 4 (Comparability) hold, but Assumption 3 (Surrogacy) does not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\text{(surrogacy-bias)} \quad \tau - \underline{\tau} = \mathbb{E}\left[\left\{\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})\right\} \cdot \frac{\rho(S_i, X_i) \cdot (1 - \rho(S_i, X_i))}{\rho(X_i) \cdot (1 - \rho(X_i))} \middle| P_i = \mathrm{E}\right].$$

(iii). Suppose Assumptions 2 (Unconfoundedness) and 3 (Surrogacy) hold, but Assumption 4 (Comparability) does not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\text{(comparability-bias)} \quad \tau - \underline{\tau} = \mathbb{E}\left[\left\{\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})\right\} \cdot \frac{\rho(S_i, X_i) - \rho(X_i)}{\rho(X_i) \cdot (1 - \rho(X_i))} \middle| P_i = \mathrm{E}\right].$$

(iv). Suppose Assumption 2 (Unconfoundedness) holds, but Assumptions 3 (Surrogacy) and 4 (Comparability) do not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\begin{aligned}\text{(total bias)} \quad \tau - \underline{\tau} = \mathbb{E}&\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right] \\ &+ \mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{\rho(S_i, X_i) - \rho(X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right].\end{aligned}$$

Remark 10. *Theorem 4(i) shows that even without Surrogacy and Comparability, we estimate a valid average causal effect as long as unconfoundedness holds. The treatment effect we estimate is the average effect of the treatment on the surrogate index – a principled aggregate of intermediate outcomes – rather than the average effect on the primary outcome. This result also shows that the interpretation does not change with the choice of estimator (using the surrogate score approach, the surrogate index approach, or the influence function). Theorem 4(ii − iv) show how violations of Comparability or Surrogacy affect the difference between what is being estimated and the average treatment effect on the primary outcome.*

Remark 11. *The bias from violations of Surrogacy (Theorem 4(ii)) consists of two factors. The first factor is small if the treatment does not explain much of the variation in Y_i and therefore $\mu(s,1,x,\mathrm{E})$ and $\mu(s,0,x,\mathrm{E})$ are close. The second factor is small if the surrogate explains a large share of the variation in W_i, so that the surrogate score is close to zero or one and therefore $\mathbb{E}[\rho(S_i,X_i)\cdot(1-\rho(S_i,X_i))]$ is close to zero.*

Remark 12. *The bias from violations of Comparability (Theorem 4(iii)) also consists of two factors. The first is the difference between the surrogacy index $\mu(s,x,\mathrm{O})$ and its counterpart in the experimental sample, $\mu(s,x,\mathrm{E})$. The second factor depends on the deviation between the surrogacy score and the propensity score, $\rho(S_i,X_i)-\rho(X_i)$. If the treatment does not have much effect on the surrogates, violations of Comparability do not generate much bias, because the bias that comes from a combination of the effect of the treatment on the surrogates and the effect of the surrogates on the outcome, will be small in that case.*

5.2 Bounds on the Bias

In this subsection we explore bounds on the parameter of interest. We show that in general these bounds are uninformative. However, if outcomes themselves are bounded, for example, if the outcomes are binary, informative bounds can be derived. Moreover, we present bounds given assumptions on the range of violations of the Surrogacy and Comparability assumptions.

Lemma 1. *Suppose Assumptions 2 (Unconfoundedness) and 4 (Comparability) hold, but Assumption 3 (Surrogacy) does not necessarily hold. Then:*
(i) If the outcome can take on values on the whole real line, then there is no value for the average treatment effect τ that can be ruled out.
(ii) if the outcome is binary, then the average treatment effect τ is inside the interval

$$\left\{\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O})-\mu(S_i(0),X_i,\mathrm{O})\right|P_i=\mathrm{E}]+\mathbb{E}\left[\Delta_S^L(S_i,X_i)\frac{\rho(S_i,X_i)(1-\rho(S_i,X_i))}{\rho(X_i)(1-\rho(X_i))}\right|P_i=\mathrm{E}\right],$$

$$\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O})-\mu(S_i(0),X_i,\mathrm{O})\right|P_i=\mathrm{E}]+\mathbb{E}\left[\Delta_S^U(S_i,P_i)\frac{\rho(S_i,X_i)(1-\rho(S_i,X_i))}{\rho(X_i)(1-\rho(X_i))}\right|P_i=\mathrm{E}\right]\Bigg\},$$

where

$$\Delta_S^L(s,x) = -\min\left(\frac{1-\mu(s,x,\mathrm{O})}{\rho(s,x)}, \frac{\mu(s,x,\mathrm{O})}{1-\rho(s,x)}\right) \qquad \Delta_S^U(s,x) = \min\left(\frac{\mu(s,x,\mathrm{O})}{\rho(s,x)}, \frac{1-\mu(s,x,\mathrm{O})}{1-\rho(s,x)}\right),$$

and this bound is sharp.

(iii) if the direct effect of the treatment on the outcome $\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})$ *is bounded in absolute value by* c*, then the average treatment effect* τ *is inside the interval*

$$\Big\{\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O}) - \mu(S_i(0),X_i,\mathrm{O})\right| P_i = \mathrm{E}] - c\cdot\mathbb{E}\left[\frac{\rho(S_i,X_i)(1-\rho(S_i,X_i))}{\rho(X_i)(1-\rho(X_i))}\right| P_i = \mathrm{E}\Big],$$

$$\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O}) - \mu(S_i(0),X_i,\mathrm{O})\right| P_i = \mathrm{E}] + c\cdot\mathbb{E}\left[\frac{\rho(S_i,X_i)(1-\rho(S_i,X_i))}{\rho(X_i)(1-\rho(X_i))}\right| P_i = \mathrm{E}\Big]\Big\},$$

and this bound is sharp.

Remark 13. *To provide some intuition for the sharpness of the bounds, consider the surrogacy bias in Theorem 4. The bias has two factors, with the second estimable from the data. The first factor is the difference* $\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})$*. The data are not directly informative about this difference beyond the fact that the weighted average* $\rho(S_i,X_i)\mu(S_i,1,X_i,\mathrm{E}) + (1-\rho(S_i,X_i))\mu(S_i,0,X_i,\mathrm{E})$ *is equal to the estimable quantity* $\mu(S_i,X_i,\mathrm{O})$*. In the absence of any restrictions on the outcome this implies there are no restrictions on* $\mu(S_i,w,X_i,\mathrm{E})$ *or on the difference* $\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})$*, and thus not on the bias or the average treatment effect. Given restrictions on the range of the outcome this representation directly leads to upper and lower bounds on the bias and the average treatment effect.*

Lemma 2. *Suppose Assumptions 2 (Unconfoundedness) and 3 (surrogacy) hold, but Assumption 4 (Comparability) does not necessarily hold. Then:*

(i) If the outcome can take on value on the whole real line, then there is no value for the average treatment effect τ *that can be ruled out.*

(ii) if the outcome is binary, then the average treatment effect τ *is inside the interval*

$$\Big\{\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O}) - \mu(S_i(0),X_i,\mathrm{O})\right| P_i = \mathrm{E}] + \mathbb{E}\left[\Big\{\mathbf{1}_{\rho(S_i,X_i)<\rho(X_i)} - \mu(S_i,X_i,\mathrm{O})\Big\}\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right| P_i = \mathrm{E}$$

$$\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O})-\mu(S_i(0),X_i,\mathrm{O})\middle|\, P_i=\mathrm{E}\right]+\mathbb{E}\left[\left\{\mathbf{1}_{\rho(S_i,X_i)>\rho(X_i)}-\mu(S_i,X_i,\mathrm{O})\right\}\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\middle|\, P_i=\mathrm{E}\right]$$

with width

$$2\mathbb{E}\left[\mathbf{1}_{\rho(S_i,X_i)>\rho(X_i)}\cdot\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)\cdot(1-\rho(X_i))}\middle|\, P_i=\mathrm{E}\right],$$

and this bound is sharp.

(iii) if $\mu(S_i,X_i,\mathrm{E})-\mu(S_i,X_i,\mathrm{O})$ *is bounded in absolute value by* c*, then the average treatment effect* τ *is inside the interval*

$$\left\{\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O})-\mu(S_i(0),X_i,\mathrm{O})\middle|\, P_i=\mathrm{E}\right]-c\cdot\mathbb{E}\left[\frac{|\rho(S_i,X_i)-\rho(X_i)|}{\rho(X_i)\cdot(1-\rho(X_i))}\middle|\, P_i=\mathrm{E}\right],\right.$$

$$\left.\mathbb{E}\left[\mu(S_i(1),X_i,\mathrm{O})-\mu(S_i(0),X_i,\mathrm{O})\middle|\, P_i=\mathrm{E}\right]+c\cdot\mathbb{E}\left[\frac{|\rho(S_i,X_i)-\rho(X_i)|}{\rho(X_i)\cdot(1-\rho(X_i))}\middle|\, P_i=\mathrm{E}\right]\right\},$$

and this bound is sharp.

6 Estimation

In this section, we first present four estimators for the average treatment effect. The first, the surrogate index estimator, is related to previously proposed estimators with the difference that in the earlier literature the surrogate index was implicitly assumed to be known. We then discuss three new alternative estimators. The last of these new estimators is a matching estimator. Although matching estimators are generally not efficient in settings with unconfoundedness (Rubin 2006; Abadie and Imbens 2006, 2016), they are widely applied, and it is instructive to see how a matching strategy can be used here.

6.1 Surrogate Index

Suppose we estimate the surrogate index as $\hat{\mu}(s,x,\mathrm{O})$ and the propensity score as $\hat{\rho}_{\mathrm{E}}(x)$. We take an average of the surrogate index in the experimental sample for the treatment and control groups, after adjusting for the propensity score. A natural estimator, corresponding to (4.2), is the following difference of the two averages over the experimental sample:

$$\hat{\tau}^{\mathrm{E}}=\frac{1}{\sum_{i=1}^{N_{\mathrm{E}}}W_i/\hat{\rho}(X_i)}\sum_{i=1}^{N_{\mathrm{E}}}\hat{\mu}(S_i,X_i,\mathrm{O})\cdot\frac{W_i}{\hat{\rho}(X_i)} \tag{6.1}$$

$$-\frac{1}{\sum_{i=1}^{N_{\rm E}}(1-W_i)/(1-\hat{\rho}(X_i))}\sum_{i=1}^{N_{\rm E}}\hat{\mu}(S_i,X_i,{\rm O})\cdot\frac{1-W_i}{1-\hat{\rho}(X_i)}.$$

We refer to this as the surrogate index estimator. Note that compared to the representation in Theorem 1, we normalize the weights so that the weights sum up to one. This tends to improve the finite sample properties of related estimators in other settings substantially (Hirano, Imbens and Ridder 2003; Busso, DiNardo and McCrary 2014).

In the case where the estimator for the surrogate index $\mu(s,x,{\rm O})$ was based on a linear specification for the regression of the primary outcome on the intermediate outcome, $\mu(s,x,{\rm O}) = \gamma_0+\gamma_S's+\gamma_X'x$, this leads to

$$\hat{\tau}^{\rm E}=\hat{\gamma}_S'\hat{\tau}_S,$$

where $\hat{\tau}_S$ is an estimator for the average effect of the treatment on the surrogates, $\mathbb{E}\left[S_i(1)-S_i(0)\right]$. In the simplest case without pre-treatment variables and where the experimental sample is randomized, $\hat{\tau}_S=\overline{S}_1-\overline{S}_0$, where $\overline{S}_1$ and $\overline{S}_0$ are the average values of the surrogate outcomes. Here, the estimator simplifies to the difference in the estimated surrogate index in the treatment group and the control group: $\hat{\tau}^{\rm E}=\hat{\gamma}_S'(\overline{S}_1-\overline{S}_0)$. This expression is also familiar from the mediation literature (*e.g.*, Baron and Kenny 1986) and the surrogacy literature (Day and Duffy, 1996). However, we emphasize that in general, there may be interactions between the surrogates and pre-treatment variables, and in that case the linear specification need not be not adequate.

6.2 Surrogate Score Estimator

We now use the second representation for τ in the main theorem to derive an alternative estimator. Let $\hat{\rho}(x)$, $\hat{\rho}(s,x),\hat{\varphi}(s,x),\hat{\varphi}(x)$, and $\hat{\varphi}$, be estimators for $\rho(x)$, $\rho(s,x),\varphi(s,x),\varphi(x)$, and φ respectively.

The surrogate score estimator is based on averaging the following expression over the observational sample:

$$\hat{\tau}^{\rm O}=\frac{1}{\sum_{i|P_i={\rm O}}\omega_{1,i}}\sum_{i|P_i={\rm O}}Y_i\cdot\omega_{1,i}-\frac{1}{\sum_{i|P_i={\rm O}}\omega_{0,i}}\sum_{i|P_i={\rm O}}Y_i\cdot\omega_{0,i},\tag{6.2}$$

where for $w = 0, 1$ the weights are

$$\omega_{w,i} = \frac{\hat{\rho}(S_i, X_i)^w \cdot (1 - \hat{\rho}(S_i, X_i))^{1-w} \cdot \hat{\varphi}(S_i, X_i) \cdot (1 - \hat{\varphi})}{\hat{\rho}(X_i)^w \cdot (1 - \hat{\rho}(X_i))^{1-w} \cdot (1 - \hat{\varphi}(S_i, X_i)) \cdot \hat{\varphi}}. \tag{6.3}$$

6.3 Influence Function Estimator

We can also base estimation on the efficient score given in (4.4). Given estimators for the propensity score, the surrogate score, and the sampling score, we can estimate the average treatment effect as

$$\begin{aligned}\hat{\tau}^{E,O} = \sum_{i=1}^{N} \Bigg\{ & \frac{\mathbf{1}_{P_i=\mathrm{E}}}{\hat{\varphi}} \left(\frac{W_i \cdot \hat{\mu}(S_i, X_i, \mathrm{O})}{\hat{\rho}(X_i)} - \frac{(1 - W_i) \cdot \hat{\mu}(S_i, X_i, \mathrm{O})}{1 - \hat{\rho}(X_i)} \right) \\ & + \frac{\mathbf{1}_{P_i=\mathrm{E}}}{\hat{\varphi}} \left(\hat{\mu}(1, X_i) \left(1 - \frac{W_i}{\hat{\rho}(X_i)} \right) - \hat{\mu}(0, X_i) \left(1 - \frac{1 - W_i}{1 - \hat{\rho}(X_i)} \right) \right) \\ & + \frac{\mathbf{1}_{P_i=\mathrm{O}}}{1 - \hat{\varphi}} \left(\frac{\hat{\varphi}(S_i, X_i)}{1 - \hat{\varphi}(S_i, X_i)} \frac{1 - \hat{\varphi}}{\hat{\varphi}} \right) \frac{(Y_i - \hat{\mu}(S_i, X_i, \mathrm{O}))\,(\hat{\rho}(S_i, X_i) - \hat{\rho}(X_i))}{\hat{\rho}(X_i)(1 - \hat{\rho}(X_i))} \Bigg\}. \end{aligned} \tag{6.4}$$

Based on the results in Newey (1994), it follows that under standard conditions the two estimators above and the surrogate index estimator all reach the semi-parametric efficiency bound, and are first-order equivalent.

The recent literature on double robust estimation of average treatment effects under unconfoundedness (Chernozhukov et al., 2016) suggests that this estimator may have superior properties in small samples.

6.4 Double Matching Estimator

Consider unit i in the experimental sample with $X_i = x$ and $S_i = s$, and suppose this is a treated unit with $W_i = 1$. We need to find three matches for this unit. First, we need to find a unit with the opposite treatment in the same (experimental) sample. Specifically, we need to find the closest unit in the experimental sample, in terms of pre-treatment variables, among the units with $W_i = 0$. Suppose this unit is unit j, with $W_j = 0$, and the value of the pre-treatment variables for this unit are $X_j = x'$, and the surrogate outcomes are $S_j = s'$. As a result of the

matching we should have $x \approx x'$, but potentially s could be quite different from s'. Next, we need to find for each of the two units i and j a match in the observational sample. Find the unit in the observational sample closest to unit i, in terms of both pre-treatment variables and surrogates. Let i' be the index for this unit, and let the value of the outcome for this unit be $Y_{i'}$, and the values of the pre-treatment variables and surrogates $X_{i'}$ and $S_{i'}$. Now as a result of the matching $X_i \approx X_{i'}$ and $S_i \approx S_{i'}$. Finally, find the unit in the observational sample closest to unit j, in terms of both pre-treatment variables and surrogates. Let the value of the outcome for this unit be $Y_{j'}$, and the values of the pre-treatment variables and surrogates $X_{j'}$ and $S_{j'}$, with $X_j \approx X_{j'}$ and $S_j \approx S_{j'}$.

Then we combine these matches to estimate the causal effect for unit i, $Y_i(1) - Y_i(0)$, as the difference in average outcomes for the two matches from the observational sample:

$$\widehat{Y_i(1) - Y_i(0)} = Y_{i'} - Y_{j'}. \tag{6.5}$$

The matching estimator for τ would then be the average value of (6.5) over the experimental sample. The double matching estimator is then

$$\hat{\tau}^{\text{match}} = \frac{1}{N^{\text{E}}} \sum_{i:P_i\text{E}} \left\{ W_i \left(Y_{i'} - Y_{j'} \right) + \left(1 - W_i \right) \left(Y_{j'} - Y_{i'} \right) \right\}.$$

7 Application: Impacts of Job Training on Employment

In this section, we apply our method to estimate the causal effect of the Greater Avenues to Independence (GAIN) job training program on long-term labor market outcomes. GAIN was a job assistance program implemented in California in the 1980s to help welfare recipients find work (Riccio et al. 1989; Friedlander and Robins 1995; Hotz, Imbens and Klerman 2006). MDRC conducted a randomized trial to evaluate the GAIN program's employment impacts in six counties in California in the late 1980s. We focus primarily on the GAIN trial in Riverside, which was widely heralded as the program that had the largest treatment effects on earnings. The Riverside program emphasized a "jobs first" approach to re-entry into the labor force, encouraging unemployed workers to take any job they find; in contrast, other sites focused more

heavily on developing human capital through training programs (Hotz, Imbens and Klerman 2006).

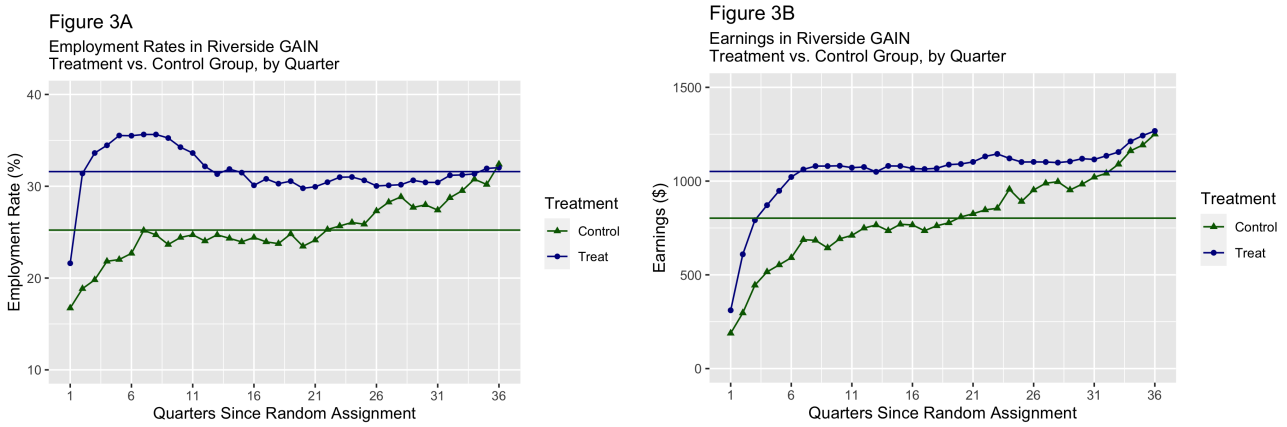
We have available long-term outcomes for the four GAIN sites, including employment, earnings, and receipt of aid over the first thirty-six quarters after random assignment. We take the average of the thirty-six employment indicators and earnings in Riverside as our primary outcomes. We then investigate whether we could have predicted the long-term impact on these outcomes using only the first T quarters of all outcomes (including employment, earnings, and aid) as surrogates, as well as using pre-treatment variables (characteristics of the individuals as well as lagged employment, earnings and aid outcomes). The Riverside data on the treatment, surrogates and pre-treatment variables play the role of our experimental sample. We use the data from the combination of the other three locations (Alameda, Los Angeles, and San Diego) as our observational sample. For the observational sample we only use the information on the surrogates, pre-treatment variables, and outcome, but not the treatment assignment, nor the indicator for the location.

We begin by presenting a brief summary of the samples. We then describe how we construct our surrogate index. Next we illustrate our theoretical results by evaluating the magnitude of the gains from using surrogate indices in terms of time and precision relative to existing experimental estimates of the program's long-term impacts in Riverside. We also show how one can validate the surrogacy assumption using intermediate outcomes and bound the degree of bias arising from potential violations of surrogacy.

7.1 The GAIN Program

The GAIN treatment was randomly assigned to welfare (Aid for Families with Dependent Children) recipients, a very low-income population. The treatment group consisted of $N_{\mathrm{E},T} = 4405$ participants, which the control group consisted of $N_{\mathrm{E},C} = 1040$ participants who were not eligible for the additional services in the GAIN program. The data we use come from the Hotz, Imbens and Klerman (2006) which followed study participants for nine years after assignment

of the treatment, measuring quarterly employment rates and earnings[2] from the Unemployment Insurance database. They found that the treatment effects of the Riverside GAIN program on employment rates and earnings were initially large, but declined over time, as shown in Figure 3A, which plots employment rates by quarter for individuals in the experimental (Riverside) treatment and control groups, and in Figure 3B, which shows the correspond results for quarterly earnings.



(a) Employment **(b)** Earnings

In Riverside, the estimated causal effects on the primary outcomes were a 6.4 (s.e. = 1.2) percentage point (pp) increase in average quarterly employment rates, and an $249 (s.e. $84) increase in average quarterly earnings, in both cases averaged over the 36 quarter post-treatment. Our question is whether these impacts could have been estimated more quickly by using short-term employment, earnings and aid receipt as surrogates.

The observational sample includes the other three locations, Alameda, Los Angeles and San Diego, for a total of $N_{\text{O}} = 13,725$ individuals.

In the online appendix Table 8 presents information on the pre-treatment variables. Clearly the two samples, Riverside and the combination of the other three locations, are substantially different prior to the intervention in terms of permanent characteristics such as ethnicity, as well

[2] All income variables were converted to 1999 dollars using cost-of-living deflators; see footnote 21 of Hotz, Imbens and Klerman (2000) for more information.

as in pre-treatment outcomes.

7.2 Three Estimators

We discuss here the estimators for the average effect of the program We wish to consider different set of surrogates, indexed by the number of periods t we want to use as surrogates. To capture this we index the surrogate for individual i, S_i^t, by the superscript t. S_i^t contains the employment indicators, earnings outcomes and aid receipt indicators for the t quarters after the intervention.

7.2.1 Surrogate Index Estimator

To construct the surrogacy index we estimate a linear regression model using least squares, for the individuals in the observational sample

$$Y_i = \beta_0 + \beta_S^\top S_i^t + \beta_X^\top X_i + \varepsilon_i. \tag{7.1}$$

The predicted value from this regression, which we denote by $\hat{Y}_i$, is our surrogate index for mean employment based on surrogates up to quarter t. We then compute this surrogate index for each of the individuals in the experimental sample and estimate the treatment effect based on the surrogate index as

$$\hat{\tau}^{\rm O} = \frac{1}{N_{{\rm E},T}} \sum_{i=1}^{N^{\rm E}} \hat{Y}_i W_i - \frac{1}{N_{{\rm E},C}} \sum_{i=1}^{N_{\rm E}} \hat{Y}_i (1 - W_i). \tag{7.2}$$

If we use the all 36 quarters of employment indicators are surrogates, then the regression of Y_i on the set of surrogates will fit perfectly, $\hat{Y}_i$ will be equal to Y_i, and the estimated effect will be identical to the original experimental estimate. The question is whether using a much more limited set of surrogates will get us close to the experimental benchmark.

7.2.2 Surrogate Score Estimator

For the surrogate score estimator we first estimate a logistic regression of the treatment indicator on the pretreatment variables and the surrogates. We specify

$$\ln\left(\frac{\rho(S_i^t, X_i)}{1-\rho(S_i^t, X_i)}\right) \equiv \ln\left(\frac{\text{pr}(W_i = 1 | S_i^t, X_i, P_i = \text{E})}{1 - \text{pr}(W_i = 1 | S_i^t, X_i, P_i = \text{E})}\right) = \alpha_0 + \alpha_S^\top S_i^t + \alpha_X^\top X_i,$$

and estimate this on the experimental (Riverside) sample.

Next we estimate the propensity score, also as a logistic regression,

$$\ln\left(\frac{\rho(X_i)}{1-\rho(X_i)}\right) \equiv \ln\left(\frac{\text{pr}(W_i=1|X_i,P_i=\text{E})}{1-\text{pr}(W_i=1|X_i,P_i=\text{E})}\right) = \delta_0 + \delta_X^\top X_i,$$

and estimate this again on the experimental (Riverside) sample. In principle the random assignment implies that the δ_X should be close to zero in this case.

Finally we estimate the comparability score

$$\ln\left(\frac{\varphi(S_i^t,X_i)}{1-\varphi(S_i^t,X_i)}\right) \equiv \ln\left(\frac{\text{pr}(P_i=\text{E}|X_i,S_i^t)}{1-\text{pr}(P_i=\text{E}|X_i,S_i^t)}\right) = \gamma_0 + \gamma_S^\top S_i^t + \gamma_X^\top X_i,$$

and estimate this on the combined observational and experimental samples.

The surrogate score estimator is based on averaging the following expression over the observational sample:

$$\hat{\tau}^{\text{O}} = \frac{1}{\sum_{i|P_i=\text{O}}\omega_{1,i}}\sum_{i|P_i=\text{O}} Y_i\cdot\omega_{1,i} - \frac{1}{\sum_{i|P_i=\text{O}}\omega_{0,i}}\sum_{i|P_i=\text{O}} Y_i\cdot\omega_{0,i}, \tag{7.3}$$

where the weights are as before in Equation (6.3).

7.2.3 Influence Function Estimator

For the influence function estimator we first estimate the surrogacy index, the surrogacy score, the propensity score, and the comparability score as before. We then plug those into the estimator in Equation (6.4).

7.3 Results

Here we discuss two sets of results. First the estimates for the average effect of the intervention on the two primary outcomes under various assumptions about the surrogates. Second, we test the Surrogacy and Comparability assumptions directly.

7.3.1 Estimation Results

As the discussion after the surrogate index estimator shows, we recover the experimental estimates if we use all 36 quarters of employment indicators as surrogates. The question is whether

Figure 4A

Varying Quarters to Construct Estimates with Covariates

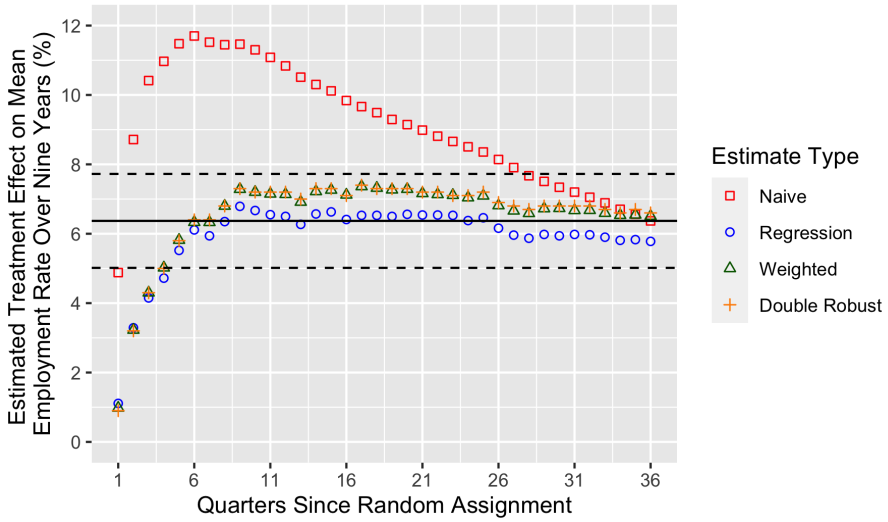
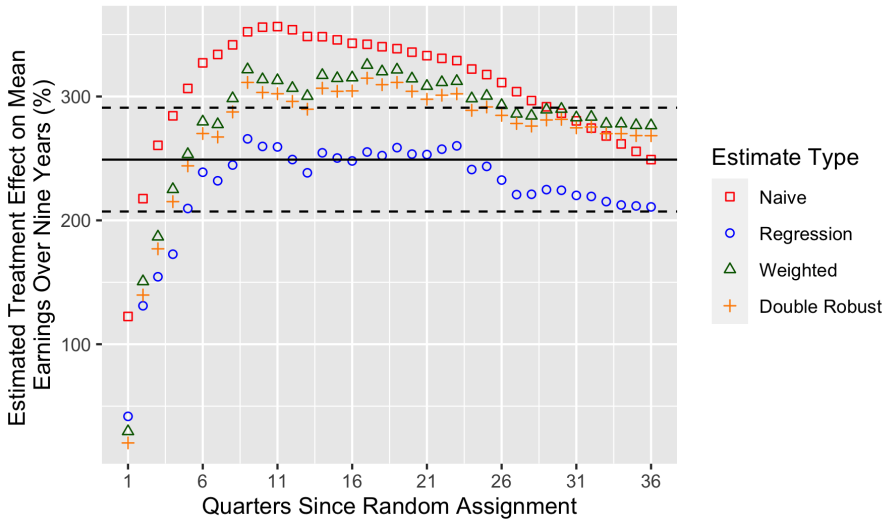


Figure 4B

Varying Quarters to Construct Estimates with Covariates



we can do approximately as well with fewer than 36 quarters of surrogates. In Figures 4A and 4B we compare the experimental estimates of the effect on the primary outcomes (0.064 for the employment outcome, and $249 for the earnings outcome) to the three sets of surrogate estimates, as a function of how many periods of surrogates we use, ranging from 1 quarter to 36 quarters. To put this in perspective we also include in these two figures what we label the "naive" estimator where we estimate the effect on the long-term outcome as the effect on the first t quarters of the outcome. In Tables 4 and 5 we report a subset of the numbers underlying these estimates with the corresponding standard errors.

Table 4: Estimates for Effect on Employment. Experimental Benchmark: 0.064 (s.e. 0.012)

t	Naive		Surrogacy Index		Surrogacy Score		Influence Function	
	est	(s.e.)	est	(s.e.)	est	(s.e.)	est	(s.e.)
1	0.049	(0.013)	0.011	(0.003)	0.010	(0.002)	0.010	(0.003)
2	0.087	(0.012)	0.033	(0.003)	0.032	(0.003)	0.033	(0.004)
3	0.104	(0.011)	0.042	(0.004)	0.043	(0.004)	0.044	(0.004)
4	0.110	(0.011)	0.047	(0.005)	0.050	(0.005)	0.052	(0.005)
5	0.115	(0.011)	0.055	(0.005)	0.058	(0.005)	0.060	(0.005)
6	0.117	(0.010)	0.061	(0.006)	0.063	(0.006)	0.065	(0.006)
12	0.108	(0.010)	0.065	(0.007)	0.071	(0.008)	0.073	(0.008)
18	0.095	(0.010)	0.065	(0.008)	0.073	(0.009)	0.075	(0.009)
24	0.085	(0.010)	0.064	(0.009)	0.070	(0.010)	0.072	(0.010)
30	0.073	(0.010)	0.059	(0.009)	0.067	(0.010)	0.070	(0.010)
36	0.064	(0.010)	0.058	(0.009)	0.065	(0.010)	0.068	(0.010)

We see that the naive estimator does very poorly. It takes more than 25 quarters before the naive estimator is within two standard errors of the experimental estimate. In contrast all three surrogate-based estimators are all within two standard errors when the surrogates include 5 quarters of outcomes, for both outcomes.

7.3.2 Validation Results and Other Supplementary Analyses

Given the data available we can also test whether using t quarters of surrogates is sufficient to satisfy Surrogacy and Comparability. To test Surrogacy we regress the primary outcome on

Table 5: Estimates for Effect on Earnings. Experimental Benchmark: $249 (s.e. $83)

t	Naive		Surrogacy Index		Surrogacy Score		Influence Function	
	est	(s.e.)	est	(s.e.)	est	(s.e.)	est	(s.e.)
1	122.4	(28.9)	41.8	(13.4)	29.6	(9.5)	31.0	(12.6)
2	217.5	(30.1)	131.1	(18.3)	150.7	(18.9)	150.4	(20.3)
3	260.6	(31.8)	154.5	(23.4)	186.8	(24.1)	187.5	(24.9)
4	284.4	(33.5)	172.7	(27.0)	225.1	(28.0)	225.1	(28.3)
5	306.5	(35.2)	209.6	(29.5)	253.4	(30.3)	254.4	(30.7)
6	327.1	(36.6)	238.8	(31.5)	279.7	(32.5)	280.6	(32.9)
12	353.9	(41.3)	249.1	(39.4)	306.8	(42.9)	308.6	(43.6)
18	340.2	(43.9)	252.3	(44.3)	320.1	(45.7)	321.8	(46.5)
24	322.2	(46.5)	241.0	(49.8)	298.3	(49.7)	300.8	(50.8)
30	286.5	(48.5)	224.3	(50.3)	289.9	(50.5)	293.0	(51.6)
36	249.1	(50.0)	210.9	(50.2)	276.6	(51.4)	279.6	(52.5)

the pre-treatment variables, the surrogates up to quarter t, and the indicator for the treatment; a finding that the treatment has an impact indicates a violation of Surrogacy. We estimate this regression using a logistic regression model, using only the data from the experimental (Riverside) sample. We report in Table 6 and 7 the results from these regressions for a number of different values for t, for the employment outcome and the earnings outcome. We report the point estimate, standard error and t-statistic. We see that point estimates for $t \leq 3$ are large and highly statistically significant. After that most of the t-statistics are less than 2, although there are some where the t-statistics are a little above 2, but the coefficient estimates are small.

We do a similar exercise for Comparability. We combine the experimental and observational samples and regress the final outcome on the surrogates, the pretreatment variables, and an indicator for the experimental sample, again using surrogates up to period t. We report the estimates on the indicator for the experimental sample, and the corresponding standard error. Here the the point estimates become smaller after $t = 12$, but the t-statistics remain large even with a substantial number of surrogate periods, indicating a violation of Surrogacy.

If we are unwilling to make the surrogacy assumption we can still calculate bounds for the

Table 6: Surrogacy and Comparability Assumption Tests for Employment Outcome

	Surrogacy Assumption			Comparability Assumption		
t	est	(s.e)	T-Stat	est	(s.e)	T-Stat
1	0.052	(0.010)	5.4	0.008	(0.005)	1.7
2	0.034	(0.009)	3.7	-0.004	(0.004)	-0.9
3	0.024	(0.009)	2.6	-0.006	(0.004)	-1.5
4	0.018	(0.009)	2.0	-0.007	(0.004)	-1.8
5	0.010	(0.008)	1.2	-0.010	(0.004)	-2.5
6	0.004	(0.008)	0.5	-0.011	(0.004)	-3.0
12	-0.004	(0.006)	-0.7	-0.015	(0.003)	-5.0
18	-0.007	(0.004)	-1.6	-0.009	(0.002)	-4.1
24	-0.005	(0.003)	-1.8	-0.005	(0.001)	-3.6
30	-0.002	(0.001)	-1.3	-0.001	(0.001)	-1.9
35	0.000	(0.000)	-2.0	0.000	(0.000)	-0.8

Table 7: Surrogacy and Comparability Assumption Tests for Earnings Outcome

	Surrogacy Assumption			Comparability Assumption		
t	est	(s.e)	T-Stat	est	(s.e)	T-Stat
1	185.6	(50.8)	3.7	-35.6	(25.3)	-1.4
2	129.7	(49.9)	2.6	-65.0	(24.5)	-2.7
3	94.3	(48.1)	2.0	-72.8	(23.5)	-3.1
4	66.3	(46.3)	1.4	-67.5	(22.4)	-3.0
5	42.2	(44.5)	1.0	-71.0	(21.5)	-3.3
6	12.6	(42.0)	0.3	-73.9	(20.5)	-3.6
12	-19.1	(31.3)	-0.6	-65.2	(15.3)	-4.2
18	-41.8	(22.0)	-1.9	-31.2	(10.8)	-2.9
24	-20.2	(13.6)	-1.5	-27.4	(6.7)	-4.1
30	-10.8	(5.9)	-1.8	-4.7	(2.8)	-1.6
35	-0.5	(1.0)	-0.5	-0.4	(0.5)	-0.8

effect on employment, using the fact that this outcome is binary. For the case where the first six quarters of post-treatment data are used as surrogates, the lower and upper bound are estimated as -0.186 and 0.124. These are not very informative, because the data now do not allow us to estimate the indirect effect of the treatment on the outcome.

Similarly, we can calculate bounds for the average effect without assuming Comparability. With six quarters of surrogates the bounds are again wide at -0.076 and 0.194 respectively. Here the fact that the treatment effect on the surrogates is strong leads to substantial sensitivity to the comparability assumption as formalized in Lemma 2.

Using the data for Riverside we can also assess the value of the Surrogacy assumption. Using the six quarters of data as surrogates, we find that the gain from knowledge of Surrogacy (the Δ in Theorem 3) is quite large. The standard error given Surrogacy, $\sqrt{\mathbb{V}_{\rm s}}$, is 0.33 times the standard error without knowledge that Surrogacy holds, $\sqrt{\mathbb{V}_{\rm ns}}$.

8 Conclusion

We develop new methods for combining intermediate outcomes to estimate the long-term impacts of treatments more rapidly and precisely. Our method requires estimating a "surrogate index" – the conditional expectation of the long-term outcome given intermediate outcomes – and then estimating the treatment effect on the surrogate index. The surrogate index can be estimated using parametric or nonparametric regression methods. We formalize conditions under which this method yields unbiased estimates, derive bounds for the degree of bias when those assumptions fail, and propose a simple out-of-sample validation approach using "hold out" intermediate outcomes. We show that surrogates can also greatly improve the precision of estimates even in settings where the treatment effect on the long-term outcome can be estimated directly, particularly when that outcome is rare or noisy.

Applying the method to analyze the impacts of the GAIN job training program in California, we find that using short-term earnings and employment rates to construct surrogate indices expedite the detection of long-term treatment effects on employment and earnings by several

years and also substantially increases precision. Furthermore, a single surrogate index accurately predicts heterogeneity in the long-term treatment effects of different types of job training programs across sites, showing that surrogate indices estimated in a given setting may be generalizable to other settings. The success of the surrogate index in this application validates the use of short-term employment outcomes as surrogates for detecting longer-term impacts of job training programs, an empirical result that can be applied when analyzing ongoing programs.

Building on this application, it would be useful to systematically establish surrogate indices that match the long-term treatment effects estimated in other experiments and quasi-experiments. Over time, this would allow researchers to collectively build a public library of surrogate indices for long-term outcomes that could be used to expedite the analysis of future interventions.

References

Abadie, Alberto, and Guido W Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *Econometrica*, 74(1): 235–267.

Abadie, Alberto, and Guido W Imbens. 2016. "Matching on the estimated propensity score." *Econometrica*, 84(2): 781–807.

Alonso, Ariel, Geert Molenberghs, Helena Geys, Marc Buyse, and Tony Vangeneugden. 2006. "A unifying approach for surrogate marker validation based on Prentice's criteria." *Statistics in medicine*, 25(2): 205–221.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association*, 91(434): 444–455.

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes." *arXiv preprint arXiv:2006.09676*.

Baron, Reuben M, and David A Kenny. 1986. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of personality and social psychology*, 51(6): 1173.

Begg, Colin B, and Denis HY Leung. 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.

Bickel, Peter J, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and Yacov Ritov. 1993. *Efficient and adaptive estimation for semiparametric models.* Vol. 4, Johns Hopkins University Press Baltimore.

Busso, Matias, John DiNardo, and Justin McCrary. 2014. "New evidence on the finite sample properties of propensity score reweighting and matching estimators." *Review of Economics and Statistics*, 96(5): 885–897.

Chen, Jiafeng, and David M Ritzwoller. 2023. "Semiparametric estimation of long-term treatment effects." *Journal of Econometrics*, 237(2): 105545.

Chen, Xiaohong, Han Hong, Alessandro Tarozzi, et al. 2008. "Semiparametric efficiency in GMM models with auxiliary data." *The Annals of Statistics*, 36(2): 808–843.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K Newey, et al. 2016. "Double machine learning for treatment and causal parameters." Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics*, 126(4): 1593–1660.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making sense of sensitivity: Extending omitted variable bias." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1): 39–67.

D'Agostino, Ralph B, Michael J Campbell, and Joel B Greenhouse. 2006. "Surrogate markers: back to the future." *Statistics in medicine*, 25(2): 181–182.

Day, NE, and SW Duffy. 1996. "Trial design based on surrogate end points – application to comparison of different breast screening frequencies." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(1): 49–60.

Fleming, Thomas R, and David L DeMets. 1996. "Surrogate end points in clinical trials: are we being misled?" *Annals of internal medicine*, 125(7): 605–613.

Frangakis, Constantine E, and Donald B Rubin. 2002. "Principal stratification in causal inference." *Biometrics*, 58(1): 21–29.

Freedman, Laurence S, Barry I Graubard, and Arthur Schatzkin. 1992. "Statistical validation of intermediate endpoints for chronic diseases." *Statistics in medicine*, 11(2): 167–178.

Friedlander, Daniel, and Philip K Robins. 1995. "Evaluating program evaluations: New evidence on commonly used nonexperimental methods." *The American Economic Review*, 923–937.

Gelman, Andrew, Gary King, and Chuanhai Liu. 1998. "Not asked and not answered: Multiple imputation for multiple surveys." *Journal of the American Statistical Association*, 93(443): 846–857.

Gilbert, Peter B, and Michael G Hudgens. 2008. "Evaluating candidate principal surrogate endpoints." *Biometrics*, 64(4): 1146–1154.

Graham, Bryan S, Cristine Campos de Xavier Pinto, and Daniel Egel. 2016. "Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST)." *Journal of Business & Economic Statistics*, 34(2): 288–301.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. "Top challenges from the first practical online controlled experiments summit." *ACM SIGKDD Explorations Newsletter*, 21(1): 20–35.

Hahn, Jinyong. 1998. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica*, 315–331.

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor economics*, 24(3): 411–482.

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

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396): 945–970.

Hotz, V Joseph, Guido Imbens, and Jacob A Klerman. 2000. "The long-term gains from GAIN: a re-analysis of the impacts of the California GAIN program."

Hotz, V Joseph, Guido W Imbens, and Jacob A Klerman. 2006. "Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program." *Journal of Labor Economics*, 24(3): 521–566.

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

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods*, 15(4): 309.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics*, 1–29.

Imbens, Guido W. 2003. "Sensitivity to exogeneity assumptions in program evaluation." *The American Economic Review, Papers and Proceedings*, 93(2): 126–132.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica*, 72(6): 1845–1857.

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

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 61: 467–476.

Joffe, Marshall M, and Tom Greene. 2009. "Related causal frameworks for surrogate outcomes." *Biometrics*, 65(2): 530–538.

Kallus, Nathan, and Xiaojie Mao. 2020. "On the role of surrogates in the efficient estimation of treatment effects with limited outcome data." *arXiv preprint arXiv:2003.12408*.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*, 604–620.

Lauritzen, Steffen L. 2004. "Discussion on causality." *Scandinavian Journal of Statistics*, 31(2): 189–193.

Little, Roderick JA, and Donald B Rubin. 2014. *Statistical analysis with missing data.* Vol. 333, John Wiley & Sons.

Little, Roderick JA, and Donald B Rubin. 2019. *Statistical analysis with missing data.* Vol. 793, Wiley.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review*, 80(2): 319–323.

Molinari, Francesca. 2020. "Microeconometrics with partial identification." *Handbook of econometrics*, 7: 355–486.

Newey, Whitney K. 1990. "Semiparametric efficiency bounds." *Journal of applied econometrics*, 5(2): 99–135.

Newey, Whitney K. 1994. "The asymptotic variance of semiparametric estimators." *Econometrica: Journal of the Econometric Society*, 1349–1382.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference.* New York, NY, USA:Cambridge University Press.

Pearl, Judea, and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science*, 29(4): 579–595.

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

Qu, Yongming, and Michael Case. 2006. "Quantifying the indirect treatment effect via surrogate markers." *Statistics in medicine*, 25(2): 223–231.

Rässler, Susanne. 2004. "Data fusion: identification problems, validity, and multiple imputation." *Austrian Journal of Statistics*, 33(1&2): 153–171.

Rässler, Susanne. 2012. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches.* Vol. 168, Springer Science & Business Media.

Riccio, James, et al. 1989. "GAIN: Early Implementation Experiences and Lessons. California's Greater Avenues for Independence Program." *Memo.*

Ridder, Geert, and Robert Moffitt. 2007. "The econometrics of data combination." *Handbook of econometrics*, 6: 5469–5547.

Robins, James M, and Andrea Rotnitzky. 1995. "Semiparametric efficiency in multivariate regression models with missing data." *Journal of the American Statistical Association*, 90(429): 122–129.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 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, Paul R. 1984. "The consequences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society: Series A (General)*, 147(5): 656–666.

Rosenbaum, Paul R, and Donald B Rubin. 1983*a*. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society. Series B (Methodological)*, 212–218.

Rosenbaum, Paul R, and Donald B Rubin. 1983*b*. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66(5): 688.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika*, 63(3): 581–592.

Rubin, Donald B. 2004. *Multiple imputation for nonresponse in surveys.* Vol. 81, John Wiley & Sons.

Rubin, Donald B. 2006. *Matched sampling for causal effects.* Cambridge University Press.

Tchetgen Tchetgen, Eric J, and Ilya Shpitser. 2014. "Estimation of a Semiparametric Natural Direct Effect Model Incorporating Baseline Covariates." *Biometrika*, 101(4): 849–864.

van der Laan, Mark J, and Maya L Petersen. 2004. "Estimation of direct and indirect causal effects in longitudinal studies." *Memo.*

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction.* Oxford University Press.

Weir, Christopher J, and Rosalind J Walley. 2006. "Statistical evaluation of biomarkers as surrogate endpoints: a literature review." *Statistics in medicine*, 25(2): 183–203.

Xu, Jane, and Scott L Zeger. 2001. "The evaluation of multiple surrogate endpoints." *Biometrics*, 57(1): 81–87.

Zheng, Wenjing, and Mark J van der Laan. 2012. "Targeted maximum likelihood estimation of natural direct effects." *The international journal of biostatistics*, 8(1): 1–40.

ONLINE APPENDICES

A. Additional Table

Table 8: Summary Statistics of Covariates by Location

	Riverside ($N_{\mathrm{E}} = 5,445$)		Other Locations ($N_{\mathrm{O}} = 13,725$)		
	Mean	(Std. Dev.)	Mean	(Std. Dev.)	t-statistic
Female	0.88	(0.33)	0.88	(0.32)	0.0
Highschool Diploma	0.52	(0.50)	0.50	(0.50)	3.1
Children <5	0.16	(0.37)	0.14	(0.34)	4.6
Single	0.87	(0.34)	0.86	(0.34)	0.4
Grade 17 to 20	0.00	(0.03)	0.00	(0.07)	-5.3
Grade 16	0.01	(0.08)	0.02	(0.12)	-5.4
Grade 13 to 15	0.11	(0.31)	0.12	(0.33)	-3.5
Grade 12	0.36	(0.48)	0.33	(0.47)	4.0
Grade 9 to 11	0.40	(0.49)	0.34	(0.47)	7.5
White	0.52	(0.50)	0.30	(0.46)	27.4
Hispanic	0.27	(0.45)	0.26	(0.44)	1.9
Black	0.16	(0.36)	0.34	(0.47)	-28.6
Age	33.64	(8.20)	35.39	(8.81)	-13.1
Lagged Aid for t = 1 Quarter	0.77	(0.42)	0.84	(0.37)	-9.8
Lagged Aid for t = 2 Quarter	0.65	(0.48)	0.77	(0.42)	-16.0
Lagged Aid for t = 3 Quarter	0.64	(0.48)	0.76	(0.43)	-16.3
Lagged Aid for t = 4 Quarter	0.63	(0.48)	0.75	(0.43)	-15.6
Lagged Earnings for t = 1 Quarter	453	(1405)	437	(1283)	0.7
Lagged Earnings for t = 2 Quarter	575	(1553)	510	(1433)	2.6
Lagged Earnings for t = 3 Quarter	598	(1601)	543	(1492)	2.2
Lagged Earnings for t = 4 Quarter	613	(1602)	571	(1582)	1.7
Lagged Earnings for t = 5 Quarter	666	(1701)	580	(1619)	3.2
Lagged Earnings for t = 6 Quarter	698	(1761)	580	(1587)	4.3
Lagged Earnings for t = 7 Quarter	709	(1789)	579	(1630)	4.6
Lagged Earnings for t = 8 Quarter	726	(1839)	567	(1631)	5.6
Lagged Earnings for t = 9 Quarter	719	(1828)	571	(1656)	5.2
Lagged Earnings for t = 10 Quarter	730	(1815)	573	(1663)	5.5

B. Related Literature

Critical Assumptions in the Mediation Literature and their Relation to Surrogacy

In the mediation literature (*e.g.,* Baron and Kenny 1986; VanderWeele 2015), the intermediate outcome that we refer to here as the surrogate S_i is called a mediator. To emphasize its

role as a causal variable in the mediation literature, we expand the notation and consider potential outcomes $Y_i(w, s)$ that are indexed by the treatment and the surrogate. (In terms of these potential outcomes the original potential outcomes defined in the previous section, $Y_i(w)$, indexed only by the treatment W_i, equals $Y_i(w) = Y_i(w, S_i(w))$, for $w \in \mathbb{W}$.) In the setting considered in the mediation literature, we observe the quadruple $(Y_i, S_i, W_i, X_i, P_i)$ for all units in the sample and so there is not necessarily a distinction between the experimental sample and the observational sample. To capture that we focus in this section on the case where we only have the experimental sample, $P_i = \text{E}$, and where we observe the primary outcome Y_i for this sample.

The focus of the mediation literature is on decomposing the causal effect of the treatment on the outcome into a direct effect that involves comparing potential outcomes where the surrogate remains fixed, and an indirect effect that passes through the mediator/surrogate. Three key estimands are the average *total effect,*

$$\tau^{\text{total}} \equiv \mathbb{E}\left[Y_i(1, S_i(1)) - Y_i(0, S_i(0))\right],$$

the average *natural indirect effect,* where we fix the treatment at $w = 1$, but change the surrogate from $S_i(0)$ to $S_i(1)$,

$$\tau^{\text{nie}} \equiv \mathbb{E}\left[Y_i(1, S_i(1)) - Y_i(1, S_i(0))\right],$$

and the average *natural direct effect*, where we fix the surrogate at $S_i(0)$ and change the treatment from $W_i = 0$ to $W_i = 1$:

$$\tau^{\text{nde}} \equiv \mathbb{E}\left[Y_i(1, S_i(0)) - Y_i(0, S_i(0))\right],$$

with the latter two adding up to the first: $\tau^{\text{total}} = \tau^{\text{nie}} + \tau^{\text{nde}}$.

These effects are identified in the mediation literature using assumptions similar to Assumptions 2 and 3. The first assumption in the mediation framework is a reformulation of the unconfoundedness assumption, Assumption 2. It rules out the presence of unmeasured confounders between the treatment and the surrogate, and between the treatment and the outcome.

Assumption 5. (Unconfounded Treatment Assignment / Strong Ignorability)
(*i*) $W_i \perp\!\!\!\perp \Big(S_i(0), S_i(1), Y_i(0, S_i(0)), Y_i(1, S_i(1))\Big) \;\Big|\; X_i, P_i = \mathrm{E}$,
(*ii*) $0 < \rho(x) < 1$ for all $x \in \mathbb{X}$.

The second assumption typically made in the mediation literature is another unconfoundedness assumption that rules out the presence of unobserved confounders between the surrogate and the outcome, conditional on the treatment.

Assumption 6.

$$S_i \perp\!\!\!\perp \Big(Y_i(W_i, s)_{s\in\mathbb{S}}\Big) \;\Big|\; W_i, X_i, P_i.$$

This assumption implies that comparisons of primary outcomes for units with different values for the surrogates but identical values for the treatment and pre-treatment variables can be given a causal interpretation.

To make the link to the surrogacy literature we need to add one key assumption that is not commonly made in the mediation literature. This assumption rules out any direct effect of the treatment on the outcome, allowing only for an indirect effect through the surrogate.

Assumption 7. *For all i,* $w, w' \in \mathbb{W}, s \in \mathbb{S}$,

$$Y_i(w, s) = Y_i(w', s).$$

This assumption is similar to the exclusion restriction in instrumental variables settings, *e.g.,* Imbens and Angrist (1994); Angrist, Imbens and Rubin (1996). In combination with the previous assumption this implies that we can give comparisons in the primary outcome between units with different values for the surrogates but the same values for pre-treatment variables a causal interpretation, without knowing the treatment status.

The following proposition links the surrogacy and mediation assumptions.

Proposition 3. *Suppose Assumptions 5-7 hold. Then Assumptions 2 and 3 hold.*

This connection highlights that at the heart of the surrogacy assumption is a causal relation between the surrogate and the primary outcome that mediates the causal effect of the treatment on the outcome.

Surrogacy and Comparability from a Missing Data Perspective

From a missing data perspective, Surrogacy and Comparability have parallels to the missingness at random (MAR) assumption common in the missing data literature (Rubin 1976; Little and Rubin 2019), and specifically the literature on combining samples with different sets of variables, (Ridder and Moffitt 2007; Gelman, King and Liu 1998; Rässler 2004; Graham, Pinto and Egel 2016). In particular (Rässler, 2012) focuses on a missing data structure closely related to ours.

In our two sample setting, we can think of the complete data as the quintuple $(Y_i, S_i, W_i, X_i, P_i)$. Here, we view the sample as randomly drawn from a large population, so that we view P_i as a stochastic missing data indicator. For the units in the sample we observe the incomplete data $(\mathbf{1}_{P_i=\mathrm{O}}Y_i, S_i, X_i, \mathbf{1}_{P_i=\mathrm{E}}W_i, P_i)$, where for units with $P_i = \mathrm{O}$ the treatment indicator W_i is missing, and for units with $P_i = \mathrm{E}$ the outcome Y_i is missing. Now consider the following assumption.

Assumption 8. (Augmented Missing At Random Assumption)
Conditional on (S_i, X_i), the three variables P_i, Y_i and W_i are jointly independent:

$$P_i \perp\!\!\!\perp Y_i \perp\!\!\!\perp W_i \,\Big|\, S_i, X_i.$$

This is slightly different from a standard MAR assumption in (Rubin, 1976) where one would assume $P_i \perp\!\!\!\perp Y_i | S_i, X_i$ and/or $P_i \perp\!\!\!\perp W_i | S_i, X_i$. We need the stronger assumption to incorporate surrogacy, as the following proposition shows.

Proposition 4. (Missing Data Model)
(*i*) *Assumption 8 implies Assumption 3 (Surrogacy)*

$$Y_i \perp\!\!\!\perp W_i \,\Big|\, S_i, X_i,$$

and Assumption 4 (Comparability)

$$P_i \perp\!\!\!\perp Y_i \,\Big|\, S_i, X_i.$$

(*ii*) *Assumption 8 has no testable implications.*

Note that even after we have dealt with the missing Y_i and missing W_i problems, we still have the missing potential outcomes, which is why we also need the unconfoundedness assumption.

C. Proofs

Proof of Proposition 1:

$$\begin{aligned}
&\text{pr}\left(W_i = 1 | Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right) = \mathbb{E}\left[W_i |\, Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[W_i |\, Y_i = y, S_i, X_i, \rho(S_i, X_i) = r, P_i = \text{E}\right] |\, Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[W_i |\, Y_i = y, S_i, X_i, P_i = \text{E}\right] |\, Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[W_i |\, S_i, X_i, P_i = \text{E}\right] |\, Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right] \\
&\quad = \mathbb{E}\left[\rho(S_i, X_i) |\, Y_i = y, \rho(S_i, X_i) = r, P_i = \text{E}\right] = \rho(S_i, X_i),
\end{aligned}$$

which proves the result. □

Proof of Proposition 2: Part (*i*) follows directly from the definitions of $\mu(\cdot, \text{E})$ and Assumption 3. Part (*ii*) follows directly from the definitions of $\mu(\cdot, \text{E})$ and $\mu(\cdot, \text{O})$ and Assumption 4. Part (*iii*) follows from parts (*i*) and (*ii*). □

Proof of Proposition 3: We wish to show that the three conditions

$$W_i \;\perp\!\!\!\perp\; \Big(S_i(0), S_i(1), Y_i(0, S_i(0)), Y_i(1, S_i(1))\Big) \;\Big|\; X_i \tag{8.1}$$

$$S_i \;\perp\!\!\!\perp\; \Big(Y_i(W_i, s)_{s \in \mathbb{S}}\Big) \;\Big|\; X_i, W_i \tag{8.2}$$

and

$$Y_i(w, s) = Y_i(w', s) \qquad \forall\; i, w, w' \in \mathbb{W}, s \in \mathbb{S}, \tag{8.3}$$

imply

$$W_i \;\perp\!\!\!\perp\; \Big(Y_i(0), Y_i(1), S_i(0), S_i(1)\Big) \;\Big|\; X_i, \tag{8.4}$$

$$W_i \perp\!\!\!\perp Y_i \,\Big|\, S_i, X_i. \tag{8.5}$$

Note that we leave out the conditioning in $P_i = \mathrm{E}$ in the last two conditions because we are focused here on the one-sample case. Condition (8.4) follows directly from (8.1) because $Y_i(w) = Y_i(w, S_i(w))$.

Condition (8.3) implies that we can write $Y_i(s)$ without ambiguity, and by (8.1), we have $W_i \perp\!\!\!\perp Y_i(s) \,\Big|\, X_i$. By (8.2) we have $S_i \perp\!\!\!\perp Y_i(s) \,\Big|\, X_i, W_i$. Combining these implies $\Big(S_i, W_i\Big) \perp\!\!\!\perp Y_i(s) \,\Big|\, X_i$. This in turn implies $W_i \perp\!\!\!\perp Y_i(s) \,\Big|\, S_i, X_i$, which in turn implies $W_i \perp\!\!\!\perp Y_i(S_i) \,\Big|\, S_i, X_i$. This is equivalent to the condition we set out to prove, $W_i \perp\!\!\!\perp Y_i \,\Big|\, S_i, X_i$. □

Proof of Proposition 4: The first part of the Proposition is immediate. For the second part, note that we can identify from the data the distributions

$$f_{Y_i|S_i,X_i,P_i}(y|s,x,\mathrm{O}), \qquad f_{W_i|S_i,X_i,P_i}(w|s,x,\mathrm{E}), \qquad \text{and} \quad f_{P_i,S_i,X_i}(p,s,x),$$

but no other distributions. That implies that the joint distribution of $(Y_i, S_i, W_i, X_i, P_i)$ implied by $f_{Y_i|S_i,W_i,X_i,P_i}(y|s,w,x,p) = f_{Y_i|S_i,X_i,P_i}(y|s,x,\mathrm{O})$, and $f_{W_i|S_i,X_i,P_i}(w|s,x,\mathrm{O}) = f_{W_i|S_i,X_i,P_i}(w|s,x,\mathrm{E})$, for all (y,s,s,w,x,p) is consistent with the data, and it also satisfies Assumption 8. □

Proof of Theorem 1: We prove the case for $\mathbb{E}[Y_i(1)|P_i = \mathrm{E}]$, specifically

$$\begin{aligned}
\mathbb{E}[Y_i(1)|P_i = \mathrm{E}] &= \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} \,\middle|\, P_i = \mathrm{E}\right] && (8.6)\\
&= \mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{\rho(X_i) \cdot (1 - \varphi(S_i, X_i)) \cdot \varphi} \,\middle|\, P_i = \mathrm{O}\right] && (8.7)\\
&= \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{\rho(X_i) \cdot (1 - \varphi(S_i, X_i)) \cdot \varphi} \,\middle|\, P_i = \mathrm{O}\right] && (8.8)
\end{aligned}$$

The proof of $\mathbb{E}[Y_i(0)|P_i = \mathrm{E}]$ is similar. The score function representation is immediate from these equalities. We note that equality (8.6) uses Assumptions 2–4 and equalities (8.7) and (8.8) only use the overlap condition, Assumption 4(*ii*).

Consider (8.6). By Assumption 2 (unconfoundedness), it follows that

$$\mathbb{E}[Y_i(1)|P_i = \mathrm{E}] = \mathbb{E}\left[Y_i \cdot \frac{W_i}{\rho(X_i)}\middle| P_i = \mathrm{E}\right].$$

Using the law of iterated expectations, we can first condition on S_i and X_i to get

$$\mathbb{E}\left[Y_i \cdot \frac{W_i}{\rho(X_i)}\middle| P_i = \mathrm{E}\right] = \mathbb{E}\left[\mathbb{E}\left[Y_i \cdot \frac{W_i}{\rho(X_i)}\middle| S_i, X_i, P_i = \mathrm{E}\right]\middle| P_i = \mathrm{E}\right].$$

By Assumption 3 (surrogacy), we have

$$\mathbb{E}\left[\mathbb{E}\left[Y_i \cdot \frac{W_i}{\rho(X_i)}\middle| S_i, X_i, P_i = \mathrm{E}\right]\middle| P_i = \mathrm{E}\right] = \mathbb{E}\left[\mathbb{E}\left[Y_i|S_i, X_i, P_i = \mathrm{E}\right] \cdot \frac{\mathbb{E}\left[W_i|S_i, X_i, P_i = \mathrm{E}\right]}{\rho(X_i)}\middle| P_i = \mathrm{E}\right]$$

By Assumption 4 (Comparability), $\mu(s, x, \mathrm{E}) = \mu(s, x, \mathrm{O})$ so that this is equal to

$$\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\mathbb{E}\left[W_i|S_i, X_i, P_i = \mathrm{E}\right]}{\rho(X_i)}\middle| P_i = \mathrm{E}\right] = \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, (X_i)}{\rho X_i)}\middle| P_i = \mathrm{E}\right]$$

Undoing the law of iterated expectations gives us the desired equality.

Consider (8.7). By the definition of $\varphi(s, x)$, we have

$$\frac{\varphi(s, x)}{(1 - \varphi(s, x))} \cdot \frac{1 - \varphi}{\varphi} = \frac{\mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{O}\right)}$$

where the common support condition assures $1 - \varphi(s, x)$ is not zero. This leads to

$$\mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i) \cdot t(S_i, X_i) \cdot (1 - \varphi)}{\rho(X_i) \cdot (1 - t(S_i, X_i)) \cdot \varphi}\middle| P_i = \mathrm{O}\right] = \mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)} \cdot \frac{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{O}\right)}\middle| P_i = \mathrm{O}\right]$$

Again, by the law of iterated expectations, conditioning on S_i and X_i leads to

$$\mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)} \cdot \frac{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{O}\right)}\middle| P_i = \mathrm{O}\right] = \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})\frac{\rho(S_i, X_i)}{\rho(X_i)} \cdot \frac{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{O}\right)}\middle| P_i = \mathrm{O}\right]$$

Using the definition of conditional expectations, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})\frac{\rho(S_i, X_i)}{\rho(X_i)} \cdot \frac{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i, X_i|\, P_i = \mathrm{O}\right)}\middle| P_i = \mathrm{O}\right] \\
&= \int \mu(s, x, \mathrm{O})\frac{\rho(s, x)}{\rho(x)} \cdot \frac{\mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{O}\right)} \cdot \mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{O}\right) dsdx \\
&= \int \mu(s, x, \mathrm{O})\frac{\rho(s, x)}{\rho(x)}\mathrm{pr}\left(S_i = s, X_i = x|\, P_i = \mathrm{E}\right) dsdx \\
&= \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})\frac{\rho(S_i, X_i)}{\rho(X_i)}\middle| P_i = \mathrm{E}\right]
\end{aligned}$$

Consider (8.8). By the law of iterated expectations conditional on S_i and X_i, we obtain

$$\mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{\rho(X_i) \cdot (1-\varphi(S_i, X_i)) \cdot \varphi} \middle| P_i = \mathrm{O}\right] = \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{\rho(X_i) \cdot (1-\varphi(S_i, X_i)) \cdot \varphi} \middle| P_i = \mathrm{O}\right]$$

where the common support condition assures $1 - \varphi(s, x)$ is not zero. Part (iii) follows from Proposition 4, which shows that Surrogacy and Comparability have no testable implications. Standard arguments then imply that unconfoundednes does not generate any testable implications. □

Proof of Theorem 2: For Part (i), we need to calculate the variance of the Efficient Influence Function (EIF) to obtain the efficiency bound. We provide the detailed calculation for completeness.[3]

Given the EIF:

$$\psi(y, s, w, x, p) = \frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\left(\frac{w \cdot (\mu(s, x, \mathrm{O}) - \mu(1, x))}{\rho(x)} - \frac{(1-w) \cdot (\mu(s, x, \mathrm{O}) - \mu(0, x))}{1-\rho(x)}\right)$$

$$+\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\Big(\mu(1, x) - \mu(0, x) - \tau\Big)$$

$$+\frac{\mathbf{1}_{p=\mathrm{O}}}{\varphi}\left(\frac{\varphi(s, x)}{1-\varphi(s, x)}\frac{(y - \mu(s, x, \mathrm{O}))\,(\rho(s, x) - \rho(x))}{\rho(x)(1-\rho(x))}\right)$$

$$\mathbb{V} = \left[\psi(Y_i, S_i, W_i, X_i)^2\right]$$

$$= \mathbb{E}\Bigg[\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\left(\frac{W_i \cdot (\mu(S_i, X_i, \mathrm{O}) - \mu(1, x))}{\rho(X_i)} - \frac{(1-W_i) \cdot (\mu(S_i, X_i, \mathrm{O}) - \mu(0, x))}{1-\rho(X_i)}\right)\right)^2$$

$$+\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\Big(\mu(1, x) - \mu(0, x) - \tau\Big)\right)^2$$

[3]While our influence function representation coincides with Chen and Ritzwoller (2023), the variance calculation resulted in a slightly different expression.

$$+\left(\frac{\mathbf{1}_{p=\mathrm{O}}}{\varphi}\left(\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\frac{(Y_i-\mu(S_i,X_i,\mathrm{O}))\,(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}\right)\right)^2\Bigg]$$

Focusing on the first block

$$\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\left(\frac{w\cdot(\mu(s,x,\mathrm{O})-\mu(1,x))}{\rho(x)}-\frac{(1-w)\cdot(\mu(s,x,\mathrm{O})-\mu(0,x))}{1-\rho(x)}\right)\right)^2,$$

noting that $w(1-w)=0$ and hence the cross-term disappearing, we only have to take the expectation of

$$\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\frac{w\cdot(\mu(s,x,\mathrm{O})-\mu(1,x))}{\rho(x)}\right)^2 \qquad \text{and} \qquad \left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\frac{(1-w)\cdot(\mu(s,x,\mathrm{O})-\mu(0,x))}{1-\rho(x)}\right)^2$$

Note that

$$\begin{aligned}
&\mathbb{E}\Bigg[\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\frac{W_i\cdot(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))}{\rho(X_i)}\right)^2\Bigg]\\
&=\mathbb{E}\left[\frac{(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2\varphi^2}\mathbb{E}\big[\mathbf{1}_{p=\mathrm{E}}W_i|S_i,X_i\big]\right] \quad (\because \text{Tower Property})\\
&=\mathbb{E}\left[\frac{(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2\varphi^2}\varphi(S_i,X_i)\mathbb{E}\big[W_i|S_i,X_i,P_i=\mathrm{E}\big]\right]\\
&=\mathbb{E}\left[\frac{(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2\varphi^2}\varphi(S_i,X_i)\rho(S_i,X_i)\right]\\
&=\mathbb{E}\left[\frac{\varphi(S_i,X_i)\rho(S_i,X_i)}{\varphi^2\rho(X_i)^2}(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2\right]
\end{aligned}$$

Likewise, we can derive

$$\begin{aligned}
&\mathbb{E}\Bigg[\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\frac{(1-W_i)\cdot(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))}{1-\rho(X_i)}\right)^2\Bigg]\\
&=\mathbb{E}\left[\frac{\varphi(S_i,X_i)(1-\rho(S_i,X_i))}{\varphi^2(1-\rho(X_i))^2}(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2\right]
\end{aligned}$$

Collectivizing the two term yields the first block:

$$\mathbb{E}\left[\frac{\varphi(S_i,X_i)}{\varphi^2}\left(\frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2+\frac{\rho(S_i,X_i)}{\rho(X_i)^2}(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2\right)\right]$$

Next, for the second block

$$\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\Big(\mu(1,x)-\mu(0,x)-\tau\Big)$$

we can likewise derive by using the Tower Property with respect to X_i that

$$\mathbb{E}\Bigg[\left(\frac{\mathbf{1}_{p=\mathrm{E}}}{\varphi}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)\right)^2\Bigg]=\mathbb{E}\Bigg[\frac{\varphi(X_i)}{\varphi^2}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2\Bigg]$$

Finally, for the third block

$$\left(\frac{\mathbf{1}_{p=\mathrm{O}}}{\varphi}\left(\frac{\varphi(s,x)}{1-\varphi(s,x)}\frac{(y-\mu(s,x,\mathrm{O}))\,(\rho(s,x)-\rho(x))}{\rho(x)(1-\rho(x))}\right)\right)^2,$$

note that

$$\begin{aligned}
&\mathbb{E}\Bigg[\left(\frac{\mathbf{1}_{P_i=\mathrm{O}}}{\varphi}\left(\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\frac{(Y_i-\mu(S_i,X_i,\mathrm{O}))\,(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}\right)\right)^2\Bigg]\\
&=\mathbb{E}\Bigg[\frac{(\varphi(S_i,X_i))^2}{\varphi^2(1-\varphi(S_i,X_i))^2}\frac{(\rho(S_i,X_i)-\rho(X_i))^2}{((\rho(X_i)(1-\rho(X_i)))^2)^2}\mathbb{E}\Big[\mathbf{1}_{P_i=\mathrm{O}}(Y_i-\mu(S_i,X_i,\mathrm{O}))^2|S_i,X_i\Big]\Bigg]\\
&=\mathbb{E}\Bigg[\frac{(\varphi(S_i,X_i))^2}{\varphi^2(1-\varphi(S_i,X_i))^2}\frac{(\rho(S_i,X_i)-\rho(X_i))^2}{(\rho(X_i)(1-\rho(X_i)))^2}\mathbb{E}\Big[(1-\varphi(S_i,X_i))(Y_i-\mu(S_i,X_i,\mathrm{O}))^2|S_i,X_i,P_i=\mathrm{O}\Big]\Bigg]\\
&=\mathbb{E}\Bigg[\frac{(\varphi(S_i,X_i))^2}{\varphi^2(1-\varphi(S_i,X_i))^2}\frac{(\rho(S_i,X_i)-\rho(X_i))^2}{(\rho(X_i)(1-\rho(X_i)))^2}(1-\varphi(S_i,X_i))\sigma^2(S_i,X_i,\mathrm{O})\Bigg]\\
&=\mathbb{E}\Bigg[\frac{(\varphi(S_i,X_i))^2}{\varphi^2(1-\varphi(S_i,X_i))}\frac{(\rho(S_i,X_i)-\rho(X_i))^2}{(\rho(X_i)(1-\rho(X_i)))^2}\sigma^2(S_i,X_i,\mathrm{O})\Bigg]\\
&=\mathbb{E}\Bigg[\frac{1-\varphi(S_i,X_i)}{\varphi^2}\left(\left(\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})\right)\Bigg]
\end{aligned}$$

Hence, adding up the three blocks (in the order from the third to the first block) yield the desired efficiency bound:

$$\begin{aligned}
\mathbb{V}&=\mathbb{E}[\psi(Y_i,S_i,W_i,X_i,P_i)^2]\\
&=\mathbb{E}\Bigg[\frac{1-\varphi(S_i,X_i)}{\varphi^2}\left(\left(\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})\right)\\
&+\frac{\varphi(X_i)}{\varphi^2}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2\\
&+\frac{\varphi(S_i,X_i)}{\varphi^2}\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg].\\
&=\mathbb{E}\Bigg[\frac{1}{\varphi^2}\frac{\varphi(S_i,X_i)^2}{1-\varphi(S_i,X_i)}\left(\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})
\end{aligned}$$

$$+\frac{\varphi(X_i)}{\varphi^2}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2$$
$$+\frac{\varphi(S_i,X_i)}{\varphi^2}\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg].$$

For part (ii), first rewrite the variance bound, normalized by the square root of the expected size of the experimental sample, φN, instead of normalized by the total sample size N, as

$$\tilde{\mathbb{V}}=\mathbb{E}\Bigg[\frac{1}{\varphi}\frac{\varphi(S_i,X_i)^2}{1-\varphi(S_i,X_i)}\left(\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})$$
$$+\frac{\varphi(X_i)}{\varphi}\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2$$
$$+\frac{\varphi(S_i,X_i)}{\varphi}\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg].$$

Next, we re-write the bound in terms of a conditional expectation in the experimental sample, rather than as the unconditional expectation, (this implies multiplying by $\varphi/\varphi(S_i,X_i)$ or $\varphi/\varphi(X_i)$ appropriately) as

$$\tilde{\mathbb{V}}=\mathbb{E}\Bigg[\frac{\varphi(S_i,X_i)}{1-\varphi(S_i,X_i)}\left(\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\sigma^2(S_i,X_i,\mathrm{O})$$
$$+\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2$$
$$+\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg|\,P_i=\mathrm{E}\Bigg].$$

Now we consider a sequence of data generating processes, where the outcome distribution in the observational sample remains fixed, and the propensity and surrogate scores remain fixed, and only the functions $\varphi(s,x)$, $\varphi(x)$ and the scalar φ change, in such a way that $\sup_{s,x}\varphi(s,x)\to 0$. The the first term converges to zero, leaving us with

$$\bar{\mathbb{V}}=\mathbb{E}\Bigg[\Big(\mu(1,X_i)-\mu(0,X_i)-\tau\Big)^2$$
$$+\left(\frac{(1-\rho(S_i,X_i))(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2}+\frac{\rho(S_i,X_i)(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg|\,P_i=\mathrm{E}\Bigg].$$

The final step is to note that $\rho(S_i, X_i) = \mathbb{E}[W_i|S_i, X_i, P_i = \mathrm{E}]$ so we can write $\bar{\mathbb{V}}$ as

$$\bar{\mathbb{V}} = \mathbb{E}\Bigg[\Big(\mu(1,X_i) - \mu(0,X_i) - \tau\Big)^2$$

$$+\left(\frac{(1-\mathbb{E}[W_i|S_i,X_i,P_i=\mathrm{E}])(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2} + \frac{\mathbb{E}[W_i|S_i,X_i,P_i=\mathrm{E}](\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg| P_i = \mathrm{E}\Bigg]$$

$$= \mathbb{E}\Bigg[\Big(\mu(1,X_i) - \mu(0,X_i) - \tau\Big)^2$$

$$+\left(\frac{(1-W_i)(\mu(S_i,X_i,\mathrm{O})-\mu(0,X_i))^2}{(1-\rho(X_i))^2} + \frac{W_i(\mu(S_i,X_i,\mathrm{O})-\mu(1,X_i))^2}{\rho(X_i)^2}\right)\Bigg| P_i = \mathrm{E}\Bigg].$$

□

Proof of Theorem 3:

The first representation of the efficiency bound without surrogacy in part (i) of the Theorem is essentially rewriting the efficiency bound in Hahn (1998), and related results in Robins and Rotnitzky (1995); Robins, Rotnitzky and Zhao (1995). The standard version of the efficiency bound is

$$\mathbb{V} = \mathbb{E}\left[\frac{\sigma^2(1,X_i)}{\rho(X_i)} + \frac{\sigma^2(0,X_i)}{1-\rho(X_i)} + \left(\mu(1,X_i) - \mu(0,X_i) - \tau\right)^2\right].$$

The proof consists of showing that this is equal to the expression for $\mathbb{V}_{\mathrm{ns}}$ in Theorem 3:

$$\mathbb{V}_{\mathrm{ns}} = \mathbb{E}\Bigg[\sigma^2(S_i,X_i,\mathrm{E})\cdot\left(\frac{\rho(S_i,X_i)}{\rho(X_i)^2} + \frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\right)$$

$$+\frac{\rho(S_i,X_i)}{\rho(X_i)^2}\cdot(\mu(S_i,X_i,\mathrm{E})-\mu(1,X_i))^2 + \frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\cdot(\mu(S_i,X_i,\mathrm{E})-\mu(0,X_i))^2$$

$$+(\mu(1,X_i)-\mu(0,X_i)-\tau)^2\Bigg].$$

which amounts to showing the equality of

$$\mathbb{E}\left[\frac{\sigma^2(1,X_i)}{\rho(X_i)} + \frac{\sigma^2(0,X_i)}{1-\rho(X_i)}\right], \tag{8.9}$$

and

$$\mathbb{E}\Bigg[\sigma^2(S_i,X_i,\mathrm{E})\cdot\left(\frac{\rho(S_i,X_i)}{\rho(X_i)^2} + \frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\right) \tag{8.10}$$

$$+\frac{\rho(S_i, X_i)}{\rho(X_i)^2} \cdot (\mu(S_i, X_i, \mathrm{E}) - \mu(1, X_i))^2 + \frac{1 - \rho(S_i, X_i)}{(1 - \rho(X_i))^2} \cdot (\mu(S_i, X_i, \mathrm{E}) - \mu(0, X_i))^2 \Bigg].$$

By unconfoundedness

$$\sigma^2(1, x) \equiv \mathbb{V}(Y_i(1)|X_i = x) = \mathbb{V}(Y_i|W_i = 1, X_i = x),$$

where as mentioned in the main text, we implicitly condition on the sampling indicator and abstract it from the notation when it does not lead to confusion.

By iterated expectations this is equal to

$$\mathbb{E}\left[\mathbb{V}\left(Y_i|W_i = 1, S_i, X_i = x\right)\middle| W_i = 1, X_i = x\right] + \mathbb{V}\left(\mathbb{E}[Y_i|W_i = 1, S_i, X_i]\middle| W_i = 1, X_i\right).$$

By surrogacy the conditional distribution of Y_i given W_i, S_i and X_i does not vary by W_i, so this is equal to

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{V}\left(Y_i|S_i, X_i = x\right)\middle| W_i = 1, X_i = x\right] + \mathbb{V}\left(\mathbb{E}[Y_i|S_i, X_i]\middle| W_i = 1, X_i\right) \\
&\quad = \mathbb{E}\left[\sigma^2(S_i, X_i)\middle| W_i = 1, X_i = x\right] + \mathbb{V}\left(\mu(S_i, X_i)\middle| W_i = 1, X_i\right).
\end{aligned}$$

For the first term,

$$\mathbb{E}\left[\sigma^2(S_i, X_i)\middle| W_i = 1, X_i = x\right] = \mathbb{E}\left[\frac{\sigma^2(S_i, X_i)\rho(S_i, X_i)}{\rho(X_i)}\middle| X_i = x\right].$$

For the second term, note that

$$\mathbb{E}\left[\mu(S_i, X_i)\middle| W_i = 1, X_i\right] = \mathbb{E}\left[\mathbb{E}[Y_i|S_i, X_i]\middle| W_i = 1, X_i\right]$$

is by surrogacy equal to $\mathbb{E}\left[\mathbb{E}[Y_i|W_i = 1, S_i, X_i]\middle| W_i = 1, X_i\right]$, which in turn by iterated expectations is equal to $\mathbb{E}\left[Y_i\middle| W_i = 1, X_i\right] = \mu(1, X_i)$. Hence the second term is

$$\begin{aligned}
&\mathbb{V}\left(\mu(S_i, X_i)\middle| W_i = 1, X_i\right) = \mathbb{E}\left[\left(\mu(S_i, X_i) - \mu(1, X_i)\right)^2\middle| W_i = 1, X_i\right] \\
&\quad = \mathbb{E}\left[\left(\mu(S_i, X_i) - \mu(1, X_i)\right)^2 \frac{\rho(S_i, X_i)}{\rho(X_i)}\right].
\end{aligned}$$

Combining the two terms and including the denominator $\rho(X_i)$, we have

$$\mathbb{E}\left[\frac{\sigma^2(1, X_i)}{\rho(X_i)}\right] = \mathbb{E}\left[\frac{\sigma^2(S_i, X_i)\rho(S_i, X_i)}{\rho(X_i)^2}\right] + \mathbb{E}\left[\left(\mu(S_i, X_i) - \mu(1, X_i)\right)^2 \frac{\rho(S_i, X_i)}{\rho(X_i)^2}\right].$$

By the same argument

$$\mathbb{E}\left[\frac{\sigma^2(0,X_i)}{1-\rho(X_i)}\right] = \mathbb{E}\left[\frac{\sigma^2(S_i,X_i)(1-\rho(S_i,X_i))}{(1-\rho(X_i))^2}\right] + \mathbb{E}\left[(\mu(S_i,X_i)-\mu(0,X_i))^2\,\frac{1-\rho(S_i,X_i)}{(1-\rho(X_i))^2}\right]$$

Hence, adding up the two equalities above shows the desired equivalence of (8.9) and (8.10). This finishes the proof of part (i) of the theorem.

Next, for part (ii) of the theorem, we derive the efficiency bound for the case with surrogacy by first deriving the efficient influence function and then deriving its variance. To derive the efficient influence function, we follow the proof in Chen and Ritzwoller (2023) and Newey (1990), specifically the following four steps: (1) constructing the tangent space, (2) deriving the pathwise derivative of the target estimand (i.e. the ATE under surrogacy), (3) showing that the conjectured efficient influence function (EIF) lies in the tangent space, and (4) showing that the pathwise derivative of the target estimand and the conjectured EIF satisfies a key condition in Newey (1990).

First, to characterize the tangent space, considering the data density where the functions f denote the density of random variables.

$$f_{Y_i,S_i,W_i,X_i}(y,s,w,x) = f_{Y_i|S_i,X_i}(y \mid s,x) f_{S_i|W_i,X_i}(s \mid w,x) f_{W_i|X_i}(w \mid x) f_{X_i}(x).$$

We assume the data density satisfies the regularity and smoothness conditions in Definition (A.1) of Newey (1990).

Let G^ϵ be a parametric submodel parameterized by $\epsilon \in [0,1]$ where $G^{\epsilon=0} = G$ and G is the true data generating model. Let f^ϵ be the corresponding density function for the parametric submodel. Then, the score of f_ϵ is

$$\begin{aligned}
&\frac{\delta}{\delta\epsilon}\log(f^\epsilon_{Y_i,S_i,W_i,X_i}(y,s,w,x)) \\
=&\frac{\delta}{\delta\epsilon}\log(f^\epsilon_{Y_i|S_i,X_i}(y \mid s,x)) + \frac{\delta}{\delta\epsilon}\log(f^\epsilon_{S_i|W_i,X_i}(s \mid w,x)) + \frac{\delta}{\delta\epsilon}\log(f^\epsilon_{W_i|X_i}(w \mid x)) + \frac{\delta}{\delta\epsilon}\log(f^\epsilon_{X_i}(x)) \\
=&Q^\epsilon_{Y_i|S_i,X_i}(y \mid s,x) + Q^\epsilon_{S_i|W_i,X_i}(s \mid w,x) + Q^\epsilon_{W_i|X_i}(w \mid x) + Q^\epsilon_{X_i}(x).
\end{aligned}$$

We use $Q(\cdot)$'s to denote the score function, i.e. $Q(\cdot) = \frac{\delta}{\delta\epsilon}\log(f^\epsilon(\cdot))$. Evaluating the derivative

at $\epsilon = 0$ leads us to the score of the true model, i.e.,

$$Q_{Y_i,S_i,W_i,X_i}(y,s,w,x) = Q_{Y_i|S_i,X_i}(y \mid s,x) + Q_{S_i|W_i,X_i}(s \mid w,x) + Q_{W_i|X_i}(w \mid x) + Q_{X_i}(x).$$

The tangent space $\mathcal{T}$ is the mean closure of a linear combination of mean-zero, square-integrable functions $\overline{Q}_1, ..., \overline{Q}_4$ that satisfy the following conditions:

$$\begin{aligned}\mathcal{T} = \Big\{ \overline{Q}(y,s,w,x) \in \mathbb{R} \mid \overline{Q}(y,s,w,x) &= \overline{Q}_1(y,s,x) + \overline{Q}_2(s,x,w) + \overline{Q}_3(w,x) + \overline{Q}_4(x) \\ &\mathbb{E}[\overline{Q}_1(Y_i,s,x) \mid S_i = s, X_i = x] = \mathbb{E}[\overline{Q}_1(Y_i,s,x) \mid S_i = s, W_i = w, X_i = x] = 0 \\ &\mathbb{E}[\overline{Q}_2(S_i,x,w) \mid X_i = x, W_i = w] = 0, \quad \mathbb{E}[\overline{Q}_3(W_i,x) \mid X_i = x] = 0, \\ &\mathbb{E}[\overline{Q}_4(X_i)] = 0 \Big\}.\end{aligned}$$

Second, we derive the pathwise derivative of our estimand. With some abuse of the integral notation, our estimand can be written as follows:

$$\begin{aligned}\tau &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{E}[Y_i \mid S_i, X_i] \mid W_i = 1, X_i\Big]\Big] - \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{E}[Y_i \mid S_i, X_i] \mid W_i = 0, X_i\Big]\Big] \\ &= \int\int\int y f_{Y_i|S_i,X_i}(y \mid s,x) f_{S_i|W_i,X_i}(s \mid 1,x) f_{X_i}(x) dy ds dx \\ &\quad - \int\int\int y f_{Y_i|S_i,X_i}(y \mid s,x) f_{S_i|W_i,X_i}(s \mid 0,x) f_{X_i}(x) dy ds dx.\end{aligned}$$

The pathwise derivative of the estimand τ is

$$
\begin{aligned}
\frac{\delta}{\delta\epsilon}\tau &= \int\int\int y \frac{\delta}{\delta\epsilon}\left\{f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) f^{\epsilon}_{X_i}(x)\right\} dydsdx \\
&\quad - \int\int\int y \frac{\delta}{\delta\epsilon}\left\{f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) f^{\epsilon}_{X_i}(x)\right\} dydsdx \\
&= \int\int\int y \Big\{ Q^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) f^{\epsilon}_{X_i}(x) \\
&\qquad + f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) Q^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) f^{\epsilon}_{X_i}(x) \\
&\qquad + f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) Q^{\epsilon}_{X_i}(x) f^{\epsilon}_{X_i}(x) \Big\} dydsdx \\
&\quad - \int\int\int y \Big\{ Q^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) f^{\epsilon}_{X_i}(x) \\
&\qquad + f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) Q^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) f^{\epsilon}_{X_i}(x) \\
&\qquad + f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) Q^{\epsilon}_{X_i}(x) f^{\epsilon}_{X_i}(x) \Big\} dydsdx \\
&= \int\int\int y Q^{\epsilon}_{Y_i|S_i,X_i} f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{X_i}(x) \Big\{ f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) - f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx \\
&\quad + \int\int\int y f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) f^{\epsilon}_{X_i}(x) \Big\{ Q^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) \\
&\qquad\qquad - Q^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx \\
&\quad + \int\int\int y f^{\epsilon}_{Y_i|S_i,X_i}(y \mid s,x) Q^{\epsilon}_{X_i}(x) f^{\epsilon}_{X_i}(x) \Big\{ f^{\epsilon}_{S_i|W_i,X_i}(s \mid 1,x) - f^{\epsilon}_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx.
\end{aligned}
$$

The derivatives above use the chain rule from calculus and the fact that

$$
\frac{\delta}{\delta\epsilon} f^{\epsilon} = \frac{\delta}{\delta\epsilon} \log(f^{\epsilon}) f^{\epsilon} = Q^{\epsilon} f^{\epsilon}
$$

Let τ' denote evaluating the above derivative at $\epsilon = 0$, i.e.

$$
\begin{aligned}
\tau' &= \int\int\int y Q_{Y_i|S_i,X_i}(y \mid s,x) f_{Y_i|S_i,X_i}(y \mid s,x) f_{X_i}(x) \Big\{ f_{S_i|W_i,X_i}(s \mid 1,x) - f_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx \\
&\quad + \int\int\int y f_{Y_i|S_i,X_i}(y \mid s,x) f_{X_i}(x) \Big\{ Q_{S_i|W_i,X_i}(s \mid 1,x) f_{S_i|W_i,X_i}(s \mid 1,x) \\
&\qquad\qquad - Q_{S_i|W_i,X_i}(s \mid w=0,x) f_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx \\
&\quad + \int\int\int y f_{Y_i|S_i,X_i}(y \mid s,x) Q_{X_i}(x) f_{X_i}(x) \Big\{ f_{S_i|W_i,X_i}(s \mid 1,x) - f_{S_i|W_i,X_i}(s \mid 0,x) \Big\} dydsdx \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\big[Y_i Q_{Y_i|S_i,X_i}(Y_i \mid S_i, X_i) \big| S_i, X_i \big] \middle| W_i = 1, X_i \right] - \mathbb{E}\left[\mathbb{E}\big[Y_i Q_{Y_i|S_i,X_i}(Y_i \mid S_i, X_i) \big| S_i, X_i \big] \middle| W_i = 0, X_i \right] \right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i) Q_{S_i|W_i,X_i}(S_i \mid W_i = 1, X_i) \middle| W_i = 1, X_i \right] - \mathbb{E}\left[\mu(S_i, X_i) Q_{S_i|W_i,X_i}(S_i \mid W_i = 0, X_i) \middle| W_i = 0, X_i \right] \right] \\
&\quad + \mathbb{E}[Q_{X_i}(X_i)(\mu(1, X_i) - \mu(0, X_i))]
\end{aligned}
$$

Third, consider the conjectured efficient influence function (EIF).

$$\begin{aligned}\psi(Y_i, S_i, W_i, X_i) = &\frac{(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i))}{\rho(X_i)(1 - \rho(X_i))} \\ &+ \frac{W_i(\mu(S_i, X_i) - \mu(1, X_i))}{\rho(X_i)} - \frac{(1 - W_i)(\mu(S_i, X_i) - \mu(0, X_i))}{1 - \rho(X_i)} \\ &+ \mu(1, X_i) - \mu(0, X_i) - \tau\end{aligned}$$

We show that $\psi(Y_i, S_i, W_i, X_i)$ is an element of the tangent space $\mathcal{T}$ by showing that different parts of $\psi(Y_i, S_i, W_i, X_i)$ satisfies conditions for $\overline{Q}_1, \overline{Q}_2$, and $\overline{Q}_4$.

1. For $\overline{Q}_1$, we have $\mathbb{E}\left[\frac{(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}\middle| S_i = s, X_i = x\right] = 0$ by definition of $\mu(S_i, X_i)$ and $\mathbb{E}\left[\frac{(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}\middle| S_i = s, W_i = w, X_i = x\right] = 0$ by using statistical surrogacy.

2. For $\overline{Q}_2$, we have $\mathbb{E}\left[\frac{W_i(\mu(S_i,X_i)-\mu(1,X_i))}{\rho(X_i)}\middle| W_i = w, X_i = x\right] = \frac{w}{\rho(x)}(\mathbb{E}[\mu(S_i, X_i) \mid W_i = w, X_i = x] - \mu(1, x)) = 0$ for any value of w. Similarly, $\mathbb{E}\left[\frac{(1-W_i)(\mu(S_i,X_i)-\mu(0,X_i))}{1-\rho(X_i)}\middle| W_i = w, X_i = x\right] = \frac{1-w}{1-\rho(x)}\left(\mathbb{E}\left[h(S_i, X_i)\middle| W_i = w, X_i = x\right] - \mu(0, x)\right) = 0$ for any value of w.

3. For $\overline{Q}_4$, we have $\mathbb{E}[\mu(1, X_i) - \mu(0, X_i) - \tau] = 0$.

By setting $\overline{Q}_3 = 0$, we arrive at $\psi(Y_i, S_i, W_i, X_i) \in \mathcal{T}$.

Fourth, we show that τ' and $\psi(Y_i, S_i, W_i, X_i)$ satisfy the following relationship that all efficient influence functions must satisfy from Theorem 2.2 in Newey (1990):

$$\tau' = \mathbb{E}[\psi(Y_i, S_i, W_i, X_i) \cdot Q(Y_i, S_i, W_i, X_i)]. \tag{8.11}$$

We break the proof of this equality into several steps.

(a) Let us consider the part of the $\psi(Y_i, S_i, W_i, X_i)$ concerning $\frac{(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}$. We

have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{(Y_i - h(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i))}{\rho(X_i)(1-\rho(X_i))} Q_{Y_i,S_i,W_i,X_i}(Y_i, S_i, W_i, X_i)\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i)) Q_{Y_i,S_i,W_i,X_i}(Y_i, S_i, W_i, X_i)\middle| X_i\right]\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}E\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i)) Q_{Y_i|,S_i,X_i}(Y_i \mid S_i, X_i)\middle| X_i\right]\right] \\
&+\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i)) Q_{S_i|W_i,X_i}(S_i \mid W_i, X_i)\middle| X_i\right]\right] \\
&+\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i)) Q_{W_i|X_i}(W_i \mid X_i)\middle| X_i\right]\right] \\
&+\mathbb{E}\left[\frac{Q_{X_i}(X_i)}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i))\middle| X_i\right]\right]
\end{aligned}
$$

The first equality uses the law of total expectation. The second equality uses the definition of Q_{Y_i,S_i,W_i,X_i}. We consider each term separately, starting from the bottom.

For the Q_{X_i} term, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{Q_{X_i}(X_i)}{\rho(X_i)(1-\rho(X_i))}E\left[(Y_i - \mu(S_i, X_i))(\rho(S_i, X_i) - \rho(X_i))\middle| X_i\right]\right] \\
=&\mathbb{E}\left[\frac{Q_{X_i}(X_i)}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(\rho(S_i, X_i) - \rho(X_i))\mathbb{E}\big[(Y_i - \mu(S_i, X_i))\big|S_i, X_i\big]\middle| X_i\right]\right] \\
=&\mathbb{E}\left[\frac{Q_4(X_i)}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(\rho(S_i, X_i) - \rho(X_i))\cdot 0\middle| X_i\right]\right] \\
=&0.
\end{aligned}
$$

The first equality uses the law of total expectation. The second equality uses the definition of $\mu(S_i, X_i)$.

For the $Q_{W_i|X_i}$ term, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))Q_{W_i|X_i}(W_i \mid X_i)\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Q_{W_i|X_i}(W_i \mid X_i)\mathbb{E}\big[(Y_i-h(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))\big|W_i,X_i\big]\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Q_{W_i|X_i}(W_i \mid X_i)\mathbb{E}\left[(\rho(S_i,X_i)-\rho(X_i))\mathbb{E}\big[(Y_i-\mu(S_i,X_i))\big|S_i,W_i,X_i\big]\middle|W_i,X_i\right]\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Q_{W_i|X_i}(W_i \mid X_i)\mathbb{E}\left[(\rho(S_i,X_i)-\rho(X_i))(\mathbb{E}\big[Y_i \mid S_i,W_i,X_i\big]-\mu(S_i,X_i))\middle|W_i,X_i\right]\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Q_{W_i|X_i}(W_i \mid X_i)\mathbb{E}\left[(\rho(S_i,X_i)-\rho(X_i))\cdot 0\middle|W_i,X_i\right]\middle| X_i\right]\right]\\
=&0
\end{aligned}$$

The first and second equalities use the law of total expectation. The third equality is algebra. The fourth equality uses statistical surrogacy.

For the $Q_{S_i|W_i,X_i}$ term, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mathbb{E}\left[(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\middle|X_i,W_i\right]\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mathbb{E}\left[(\rho(S_i,X_i)-\rho(X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\mathbb{E}\left[Y_i-\mu(S_i,X_i)\middle|S_i,X_i,W_i\right]\middle|X_i,W_i\right]\middle| X_i\right]\right]\\
=&0
\end{aligned}$$

The first two equalities use the law of total expectation. The third equality uses statistical surrogacy.

For the $Q_{Y_i|S_i,X_i}$ term, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle| X_i\right]\right]\\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Y_i(\rho(S_i,X_i)-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle| X_i\right]\right]\\
&-\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mu(S_i,X_i)(\rho(S_i,X_i)-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle| X_i\right]\right]
\end{aligned}$$

The first term above is equal to $\mathbb{E}\left[\frac{Y_i(W_i-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i|S_i,X_i)}{\rho(X_i)(1-\rho(X_i))}\right]$ because

$$
\begin{aligned}
&\mathbb{E}\left[\frac{Y_i(W_i-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{\rho(X_i)(1-\rho(X_i))}\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Y_i(W_i-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid S_i,X_i\right]\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle|S_i,X_i\right]\mathbb{E}\left[W_i-\rho(X_i)\middle|S_i,X_i\right]\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle|S_i,X_i\right](\rho(S_i,X_i)-\rho(X_i))\right] \\
=&\mathbb{E}\left[\frac{Y_i(\rho(S_i,X_i)-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{\rho(X_i)(1-\rho(X_i))}\right]
\end{aligned}
$$

The first equality uses the law of total expectation. The second equality uses statistical surrogacy where $Y_i \perp W_i|S_i,X_i$ implies $Y_i,S_i,X_i \perp W_i,X_i|S_i,X_i$. The third equality is the definition of the surrogate score. The fourth equality uses the law of total expectation. The second term above simplifies to zero because

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mu(S_i,X_i)(\rho(S_i,X_i)-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle|X_i\right]\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mu(S_i,X_i)(\rho(S_i,X_i)-\rho(X_i))\mathbb{E}\left[Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\middle|S_i,X_i\right]\middle|X_i\right]\right] \\
=&\mathbb{E}\left[\frac{1}{\rho(X_i)(1-\rho(X_i))}\mathbb{E}\left[\mu(S_i,X_i)(\rho(S_i,X_i)-\rho(X_i))\cdot 0\middle|X_i\right]\right] \\
=&0
\end{aligned}
$$

The first equality uses the law of total expectation. The second equality uses the the mean-zero property of the score function $Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)$.

Finally, we can rewrite $\frac{Y_i(W_i-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i|S_i,X_i)}{\rho(X_i)(1-\rho(X_i))}$ as

$$
\frac{Y_i(W_i-\rho(X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{\rho(X_i)(1-\rho(X_i))} = \frac{Y_iW_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{\rho(X_i)} - \frac{Y_i(1-W_i)Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{1-\rho(X_i)}
$$

Also, in expectation, each term above equals to

$$\begin{aligned}
\mathbb{E}\Big[\frac{W_iY_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{\rho(X_i)}\Big] &= \mathbb{E}\Big[\frac{1}{\rho(X_i)}\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=1,X_i\Big]\rho(X_i)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=1,X_i\Big]\Big], \\
\mathbb{E}\Big[\frac{(1-W_i)Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)}{1-\rho(X_i)}\Big] &= \mathbb{E}\Big[\frac{1}{1-\rho(X_i)}\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=0,X_i\Big](1-\rho(X_i))\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=0,X_i\Big]\Big].
\end{aligned}$$

The first equality uses the law of total expectation and the definition of the propensity score. The second equality is algebra. Overall, we have

$$\begin{aligned}
&\mathbb{E}\Big[\frac{(Y_i-\mu(S_i,X_i))(\rho(S_i,X_i)-\rho(X_i))}{\rho(X_i)(1-\rho(X_i))}Q_{Y_i,S_i,W_i,X_i}(Y_i,S_i,W_i,X_i)\Big] \\
=&\mathbb{E}\Big[\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=1,X_i\Big]\Big] - \mathbb{E}\Big[\mathbb{E}\Big[Y_iQ_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid W_i=0,X_i\Big]\Big].
\end{aligned}$$

(b) Let's consider the part of the $\psi(Y_i,S_i,W_i,X_i)$ concerning $\frac{W_i(\mu(S_i,X_i)-\mu(1,X_i))}{\rho(X_i)}$. We have

$$\begin{aligned}
&\mathbb{E}\Big[\frac{W_i(\mu(S_i,X_i)-\mu(1,X_i))}{\rho(X_i)}Q_{Y_i,S_i,W_i,X_i}(Y_i,S_i,W_i,X_i)\Big] \\
=&\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}\mathbb{E}\Big[(\mu(S_i,X_i)-\mu(1,X_i))Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\Big|W_i,X_i\Big]\Big] \\
&+\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}\mathbb{E}\Big[(\mu(S_i,X_i)-\mu(1,X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\Big|W_i,X_i\Big]\Big] \\
&+\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}Q_{W_i|X_i}(W_i \mid X_i)\mathbb{E}\Big[\mu(S_i,X_i)-\mu(1,X_i)\Big|W_i,X_i\Big]\Big] \\
&+\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}Q_{X_i}(X_i)\mathbb{E}\Big[\mu(S_i,X_i)-\mu(1,X_i)\Big|W_i,X_i\Big]\Big] \\
=&\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}\mathbb{E}\Big[(\mu(S_i,X_i)-\mu(1,X_i))\mathbb{E}\Big[Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)\Big|S_i,W_i,X_i\Big]\Big|W_i,X_i\Big]\Big] \\
&+\mathbb{E}\Big[\frac{W_i}{\rho(X_i)}\mathbb{E}\Big[(\mu(S_i,X_i)-\mu(1,X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\Big|W_i,X_i\Big]\Big] \\
=&\mathbb{E}\Big[\frac{W_i(\mu(S_i,X_i)-\mu(1,X_i))Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)}{\rho(X_i)}\Big]
\end{aligned}$$

The first equality uses the law of total expectation. The third equality uses the relationship $\mathbb{E}[\mu(S_i,X_i)|W_i=w,X_i]=\mu(w,X_i)$. The fourth equality uses the mean-zero property of the score and the law of total expectation.

We can further simplify the above expression by noticing that

$$\mathbb{E}\Big[\frac{W_i\mu(1,X_i)Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)}{\rho(X_i)}\Big] = \mathbb{E}\Big[\frac{W_i\mu(1,X_i)}{\rho(X_i)}\mathbb{E}\Big[Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)\Big|W_i,X_i\Big]\Big] = 0$$

The first equality uses the law of total expectation. The second equality uses the mean-zero property of the score function. Also,

$$\mathbb{E}\left[\frac{W_i\mu(S_i,X_i)Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)}{\rho(X_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\mu(S_i,X_i)Q_{S_i|W_i,X_i}(S_i \mid W_i=1,X_i)\middle|W_i=1,X_i\right]\middle|X_i\right]$$

The first equality uses the law of total expectation and the definition of conditional expectation with the definition $\mathbb{E}[W_i \mid X_i] = \rho(X_i)$.

Overall, we end up with the following expression

$$\mathbb{E}\left[\frac{W_i(\mu(S_i,X_i)-\mu(1,X_i))}{\rho(X_i)}Q_{Y_i,S_i,W_i,X_i}(Y_i,S_i,W_i,X_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\mu(S_i,X_i)Q_{S_i|W_i,X_i}(S_i \mid W_i=1,X_i)\middle|X_i,W_i=1\right]\middle|X_i\right]$$

(c) Let's consider the part of the $\psi(Y_i,S_i,W_i,X_i)$ concerning $\frac{(1-W_i)(\mu(S_i,X_i)-\mu(0,X_i))}{1-\rho(X_i)}$. From the above exercise, we end up with

$$\mathbb{E}\left[\frac{(1-W_i)(\mu(S_i,X_i)-\mu(0,X_i))}{1-\rho(X_i)}Q_{Y_i,S_i,W_i,X_i}(Y_i,S_i,W_i,X_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\mu(S_i,X_i)Q_{S_i|W_i,X_i}(S_i \mid W_i=0,X_i)\middle|W_i=0,X_i\right]\middle|X_i\right]$$

(d) Let's consider the part of the $\psi(Y_i,S_i,W_i,X_i)$ concerning $\mu(1,X_i)-\mu(0,X_i)-\tau$. We have

$$\begin{aligned}
&\mathbb{E}[(\mu(1,X_i)-\mu(0,X_i)-\tau)Q_{Y_i,S_i,W_i,X_i}(Y_i,S_i,W_i,X_i)]\\
=&\mathbb{E}[(\mu(1,X_i)-\mu(0,X_i)-\tau)\mathbb{E}[Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i)+Q_{S_i|W_i,X_i}(S_i \mid W_i,X_i)+Q_{X_i}(X_i) \mid X_i]]\\
=&E[(\mu(1,X_i)-\mu(0,X_i)-\tau)Q_{X_i}(X_i)]\\
=&E[(\mu(1,X_i)-\mu(0,X_i))Q_{X_i}(X_i)]
\end{aligned}$$

The first equality uses the law of total expectation. The second equality uses the property of the score where $E[Q_{W_i|X_i}(W_i \mid X_i) \mid X_i] = 0$. The third equality uses both the law of total expectation and the property of the score where

$$\mathbb{E}[Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid X_i] = \mathbb{E}[\mathbb{E}[Q_{Y_i|S_i,X_i}(Y_i \mid S_i,X_i) \mid S_i,X_i] \mid X_i] = \mathbb{E}[0 \mid X_i] = 0$$

Combining the four steps (a)-(d) arrives at the desired equality between τ' and $\psi(Y_i,S_i,W_i,X_i)$.

Finally, note that the $\mathbb{V}_s$ is obtained by calculating the variance of the EIF (already written in Theorem 3):[4]

[4]We henceforth explicitly show the conditioning $P_i = \mathrm{E}$ to be consistent with the notation in our Theorem statement.

$$\psi(Y_i, S_i, W_i, X_i, P_i) = \frac{(Y_i - \mu(S_i, X_i, \mathrm{E}))(\rho(S_i, X_i) - \rho(X_i))}{\rho(X_i)(1-\rho(X_i))} + \frac{W_i(\mu(S_i, X_i, \mathrm{E}) - \mu(1, X_i))}{\rho(X_i)}$$

$$-\frac{(1-W_i)(\mu(S_i, X_i, \mathrm{E}) - \mu(0, X_i))}{1-\rho(X_i)} + \mu(1, X_i) - \mu(0, X_i) - \tau,$$

i.e.,

$$\mathbb{V}_{\mathrm{s}} = \left[\psi(Y_i, S_i, W_i, X_i, P_i)^2\right] = \mathbb{E}\Bigg[\left(\frac{(Y_i - \mu(S_i, X_i, \mathrm{E}))(\rho(S_i, X_i) - \rho(X_i))}{\rho(X_i)(1-\rho(X_i))}\right)^2 + \left(\frac{W_i(\mu(S_i, X_i, \mathrm{E}) - \mu(1, X_i))}{\rho(X_i)}\right)^2$$

$$+\left(\frac{(1-W_i)(\mu(S_i, X_i, \mathrm{E}) - \mu(0, X_i))}{1-\rho(X_i)}\right)^2 + (\mu(1, X_i) - \mu(0, X_i) - \tau)^2\Bigg]$$

$$= \mathbb{E}\left[\sigma^2(S_i, X_i, \mathrm{E})\left(\frac{\rho(S_i, X_i) - \rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2 + (\mu(1, X_i) - \mu(0, X_i) - \tau)^2\right.$$

$$\left.+\frac{W_i}{\rho(X_i)^2}(\mu(S_i, X_i, \mathrm{E}) - \mu(1, X_i))^2 + \frac{1-W_i}{(1-\rho(X_i))^2}(\mu(S_i, X_i, \mathrm{E}) - \mu(0, X_i))^2\right]$$

$$= \mathbb{E}\left[\sigma^2(S_i, X_i, \mathrm{E})\left(\frac{\rho(S_i, X_i) - \rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2 + (\mu(1, X_i) - \mu(0, X_i) - \tau)^2\right.$$

$$\left.+\frac{\rho(S_i, X_i)}{\rho(X_i)^2}(\mu(S_i, X_i, \mathrm{E}) - \mu(1, X_i))^2 + \frac{1-\rho(S_i, X_i)}{(1-\rho(X_i))^2}(\mu(S_i, X_i, \mathrm{E}) - \mu(0, X_i))^2\right]$$

by the law of iterated expectations, and hence we have that

$$\Delta = \mathbb{V}_{\mathrm{ns}} - \mathbb{V}_{\mathrm{s}} = \mathbb{E}\left[\sigma^2(S_i, X_i, \mathrm{E})\left(\frac{\rho(S_i, X_i)}{\rho(X_i)^2} + \frac{1-\rho(S_i, X_i)}{(1-\rho(X_i))^2} - \left(\frac{\rho(S_i, X_i) - \rho(X_i)}{\rho(X_i)(1-\rho(X_i))}\right)^2\right)\right]$$

$$= \mathbb{E}\left[\sigma^2(S_i, X_i, \mathrm{E})\frac{\rho(S_i, X_i)(1-\rho(S_i, X_i))}{\rho(X_i)^2(1-\rho(X_i))^2}\right]$$

□

Proof of Theorem 4: Consider part (i). By the law of iterated expectations conditional on S_i and X_i, we have

$$\tau^{\mathrm{E}} \equiv \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-W_i}{1-\rho(X_i)}\middle| P_i = \mathrm{E}\right]$$

$$= \mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-\rho(S_i, X_i)}{1-\rho(X_i)}\middle| P_i = \mathrm{E}\right]$$

By the proof of (8.7) in Theorem 1 where we don't use Surrogacy or Comparability, we get

$$\begin{aligned}\tau^{\mathrm{O}} \equiv &\mathbb{E}\left[Y_i \cdot \frac{\rho(S_i, X_i) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{\rho(X_i) \cdot (1-\varphi(S_i, X_i)) \cdot \varphi} - Y_i \cdot \frac{(1-\rho(S_i, X_i)) \cdot \varphi(S_i, X_i) \cdot (1-\varphi)}{(1-\rho(X_i)) \cdot (1-\varphi(S_i, X_i)) \cdot \varphi} \middle| P_i = \mathrm{O}\right] \\ = &\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-\rho(S_i, X_i)}{1-\rho(X_i)} \middle| P_i = \mathrm{E}\right]\end{aligned}$$

The second equality in $\tau^{\mathrm{E}} = \tau^{\mathrm{O}} = \tau^{\mathrm{E,O}}$ is immediate based on only the law of iterated expectations. Finally, by the law of iterated expectations conditional on X_i, we have

$$\begin{aligned}&\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-W_i}{1-\rho(X_i)} \middle| P_i = \mathrm{E}\right] \\ = &\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-W_i}{1-\rho(X_i)} \middle| X_i, P_i = \mathrm{E}\right] \middle| P_i = \mathrm{E}\right]\end{aligned}$$

By Assumption 2 (unconfoundedness), we have

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i, X_i, \mathrm{O}) \cdot \frac{1-W_i}{1-\rho(X_i)} \middle| X_i, P_i = \mathrm{E}\right] \middle| P_i = \mathrm{E}\right] \\ = &\mathbb{E}\left[\mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) \cdot \frac{W_i}{\rho(X_i)} - \mu(S_i(0), X_i, \mathrm{O}) \cdot \frac{1-W_i}{1-\rho(X_i)} \middle| X_i, P_i = \mathrm{E}\right] \middle| P_i = \mathrm{E}\right] \\ = &\mathbb{E}\left[\mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) \mid X_i, P_i = \mathrm{E}\right] - \mathbb{E}\left[\mu(S_i(0), X_i, \mathrm{O}) \mid X_i, P_i = \mathrm{E}\right] \middle| P_i = \mathrm{E}\right]\end{aligned}$$

Undoing the law of iterated expectations give the desired result.

For parts (ii)-(iv), we prove (iv) first. By Assumption 2 (unconfoundedness), we have

$$\tau = \mathbb{E}\left[\mathbb{E}\left[Y_i | W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i | W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right].$$

By iterated expectations, this is equal to

$$\begin{aligned}\tau = &\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y_i | S_i, W_i = 1, X_i, P_i = \mathrm{E}\right] | W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\ &- \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y_i | S_i, W_i = 0, X_i, P_i = \mathrm{E}\right] | W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\ = &\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E}) | W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E}) | W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right]\end{aligned}$$

Thus, we have

$$\begin{aligned}&\tau - \mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) - \mu(S_i(0), X_i, \mathrm{O}) \mid P_i = \mathrm{E}\right] \\ = &\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E}) | W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E}) | W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\ &- \left\{\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \mid W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \mid W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right]\right\}\end{aligned}$$

We add and subtract

$$\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E}) | W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E}) | W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right]$$

to get

$$
\begin{aligned}
&\tau - \mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) - \mu(S_i(0), X_i, \mathrm{O}) \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] + \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&\quad - \{\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \mid W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \mid W_i = 0, X_i P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right]\}
\end{aligned}
$$

Rearranging the terms, we have

$$\tau - \mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) - \mu(S_i(0), X_i, \mathrm{O}) \mid P_i = \mathrm{E}\right] \tag{8.12}$$

$$=\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \tag{8.13}$$

$$- \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] + \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \tag{8.14}$$

$$+ \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \tag{8.15}$$

$$+ \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \tag{8.16}$$

Next, by the definition of expectations,

$$
\begin{aligned}
\mu(s, x, \mathrm{E}) =&\mathbb{E}[Y_i|S_i = s, X_i = x, P_i = \mathrm{E}] \\
=&\mathbb{E}[Y_i|S_i = s, W_i = 1, X_i = x, P_i = \mathrm{E}] \cdot \mathrm{pr}(W_i = 1|S_i = s, X_i = x, P_i = \mathrm{E}) \\
&+ \mathbb{E}[Y_i|S_i = s, W_i = 0, X_i = x, P_i = \mathrm{E}] \cdot \mathrm{pr}(W_i = 0|S_i = s, X_i = x, P_i = \mathrm{E}) \\
=&\mu(s, 1, x, \mathrm{E}) \cdot \rho(s, x) + \mu(s, 0, x, \mathrm{E}) \cdot (1 - \rho(s, x))
\end{aligned}
$$

Use this to write (8.13) as

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] P_i = \mathrm{E}\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E}) \cdot \rho(S_i, X_i) + \mu(S_i, 0, X_i, \mathrm{E}) \cdot (1 - \rho(S_i, X_i))|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot (1 - \rho(S_i, X_i))|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{\rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i))\rho(S_i, X_i)}{\rho(X_i)} \mid P_i = \mathrm{E}\right]
\end{aligned}
$$

Using the same argument we can write (8.14) as

$$
\begin{aligned}
&- \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] + \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&- \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 0, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, 1, X_i, \mathrm{E}) \cdot \rho(S_i, X_i) + \mu(S_i, 0, X_i, \mathrm{E}) \cdot (1 - \rho(S_i, X_i))|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \rho(S_i, X_i)|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{1 - \rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{1 - \rho(X_i)} \mid P_i = \mathrm{E}\right]
\end{aligned}
$$

Combining the results for (8.13) and (8.14) leads to

$$
\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right]
$$

Collecting the last two terms, (8.15) and (8.16), we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E})|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E}) \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{O}) \cdot \frac{1 - \rho(S_i, X_i)}{1 - \rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[\mu(S_i, X_i, \mathrm{E}) \cdot \frac{1 - \rho(S_i, X_i)}{1 - \rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{\rho(S_i, X_i)}{\rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
&- \mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{1 - \rho(S_i, X_i)}{1 - \rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{\rho(S_i, X_i) - \rho(X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)}|X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{\rho(S_i, X_i) - \rho(X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right]
\end{aligned}
$$

Combining the terms together, we obtain the expression in (iv)

$$
\begin{aligned}
&\tau - \mathbb{E}\left[\mu(S_i(1), X_i, \mathrm{O}) - \mu(S_i(0), X_i, \mathrm{O}) \mid P_i = \mathrm{E}\right] \\
=&\mathbb{E}\left[(\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E})) \cdot \frac{(1 - \rho(S_i, X_i)) \cdot \rho(S_i, X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right] \\
&+ \mathbb{E}\left[(\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O})) \cdot \frac{\rho(S_i, X_i) - \rho(X_i)}{(1 - \rho(X_i)) \cdot \rho(X_i)} \mid P_i = \mathrm{E}\right]
\end{aligned}
$$

Finally for part (ii), under Assumption 4 (Comparability), but not Assumption 3 (Surrogacy), $\mu(S_i, X_i, \mathrm{E}) - \mu(S_i, X_i, \mathrm{O}) = 0$ and the result is immediate from (iv). For part (iii), under Assumption 3 (Surrogacy), but not Assumption 4 (Comparability), $\mu(S_i, 1, X_i, \mathrm{E}) - \mu(S_i, 0, X_i, \mathrm{E}) = 0$ and the result is immediate from (iv). □

Proof of Lemma 1 We can identify, given overlap, the surrogate score $\rho(s,x)$, the propensity score $\rho(X)$, the surrogate index $\mu(s,x,\mathrm{O})$, and the joint distribution of (S_i, X_i, P_i). This implies that to derive upper and lower bounds we just need to derive upper and lower bounds for the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$ for each value of (s,x) and then integrate these bounds. We will demonstrate the sharpness of these bounds by showing that there exist data distributions consistent with all assumptions such that these bounds are achieved.

Part (i): By Theorem 4 the surrogacy bias can be characterized as

$$\text{surrogacy-bias} = \mathbb{E}\left[\left\{\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})\right\} \cdot \frac{\rho(S_i,X_i)\cdot(1-\rho(S_i,X_i))}{\rho(X_i)\cdot(1-\rho(X_i))}\,\middle|\, P_i = \mathrm{E}\right].$$

The data are not directly informative about the two conditional expectation $\mu(s,w,x,\mathrm{E})$ (because we do not observe the outcome in the experimental sample) beyond their relation to the surrogacy index:

$$\mu(s,x,\mathrm{O}) = \rho(s,x)\mu(s,1,x,\mathrm{E}) + (1-\rho(s,x))\mu(s,0,x,\mathrm{E}), \qquad \forall s,x.$$

This implies the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$ can be written as

$$\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E}) = \frac{\mu(s,x,\mathrm{O})}{\rho(s,x)} - \frac{\mu(s,0,x,\mathrm{E})}{\rho(s,x)}.$$

Fixing $\mu(s,x,\mathrm{O})$, $\rho(s,x)$, and $\mu(s,0,x,\mathrm{E})$ this places no restrictions on the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$ and thus no restrictions on the bias, and therefore any value for the treatment effect on the whole real line is consistent with the data in the absence of surrogacy.

Part (ii): If the outcome is binary, then some values can be ruled out. Because $\mu(s,w,x,\mathrm{E})$ is the conditional expectation of the outcome given some conditioning variables, it obviously must be inside the interval $[0,1]$, and both $\mu(s,1,x,\mathrm{E})$ and $\mu(s,0,x,\mathrm{E})$ must lie inside the interval $[0,1]$. This directly implies that $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E}) \in [-1,1]$. However, we can sharpen these bounds exploiting the fact that $\mu(s,x,\mathrm{O}) = \rho(s,x)\mu(s,1,x,\mathrm{E}) + (1-\rho(s,x))\mu(s,0,x,\mathrm{E})$. This implies that

$$\mu(s,1,x,\mathrm{E}) = \frac{\mu(s,x,\mathrm{O}) - \mu(s,0,x,\mathrm{E})(1-\rho(s,x))}{\rho(s,x)}. \tag{8.17}$$

First consider the upper bound on $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$. The question is what the pairs of values $(\mu(s,1,x,\mathrm{E}), \mu(s,0,x,\mathrm{E}))$ are that both lie inside $[0,1]$, such that $\mu(s,x,\mathrm{O}) = \rho(s,x)\mu(s,1,x,\mathrm{E}) + (1-\rho(s,x))\mu(s,0,x,\mathrm{E})$ for given $\mu(s,x,\mathrm{O})$ and $\rho(s,x)$, and that maximize the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$. There are two possibilities. Either $\mu(s,x,\mathrm{O}) \geq \rho(s,x)$ or $\mu(s,x,\mathrm{O}) < \rho(s,x)$.

If $\mu(s,x,\mathrm{O}) \geq \rho(s,x)$, then the smallest value for $\mu(s,0,x,\mathrm{E})$ such that the value for $\mu(s,x,\mathrm{E})$ implied by (8.17) is less than or equal to one is $\mu(s,0,x,\mathrm{E}) = (\mu(s,x,\mathrm{O}) - \rho(s,x))/(1 - \rho(s,x))$. This value has to be less than one by the assumption that there is a pair of values $(\mu(s,0,x,\mathrm{E}), \mu(s,1,x,\mathrm{E}))$ that satisfies (8.17). In this case upper bound for the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$ is equal to $(1-\mu(s,x,\mathrm{O}))/(1-\rho(s,x))$. If $\mu(s,x,\mathrm{O}) \leq \rho(s,x)$, then the largest value for $\mu(s,1,x,\mathrm{E})$ such that $\mu(s,0,x,\mathrm{E})$ is nonnegative is $\mu(s,x,\mathrm{O})/\rho(s,x)$. In that case the upper bound for the difference $\mu(s,1,x,\mathrm{E}) - \mu(s,0,x,\mathrm{E})$ is equal to $\mu(s,x,\mathrm{O})/\rho(s,x)$.

In summary, to demonstrate sharpness, consider the following data distributions:

If $\mu(s,x,\mathrm{O}) \geq \rho(s,x)$, set $\mu(s,0,x,\mathrm{E}) = \frac{\mu(s,x,\mathrm{O})-\rho(s,x)}{1-\rho(s,x)}$ and $\mu(s,1,x,\mathrm{E}) = 1$.

If $\mu(s,x,\mathrm{O}) < \rho(s,x)$, set $\mu(s,0,x,\mathrm{E}) = 0$ and $\mu(s,1,x,\mathrm{E}) = \frac{\mu(s,x,\mathrm{O})}{\rho(s,x)}$

In both cases, these distributions are admissible under our assumptions, and also achieve the bounds, demonstrating that the bounds are sharp.

Therefore, the sharp upper bound is

$$\Delta_S^U(s,x) = \begin{cases} (1-\mu(s,x,\mathrm{O}))/(1-\rho(s,x)) & \text{if} \quad \mu(s,x,\mathrm{O}) \geq \rho(s,x) \\ \mu(s,x,\mathrm{O})/\rho(s,x) & \text{if} \quad \mu(s,x,\mathrm{O}) \leq \rho(s,x), \end{cases}$$

$$= \min\left(\frac{\mu(s,x,\mathrm{O})}{\rho(s,x)}, \frac{1-\mu(s,x,\mathrm{O})}{1-\rho(s,x)}\right).$$

The proof for the lower bound follows the same argument.

Part (*iii*):

$$\text{surrogacy-bias} = \mathbb{E}\left[\left\{\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})\right\} \cdot \frac{\rho(S_i,X_i)\cdot(1-\rho(S_i,X_i))}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right]$$

$$\leq \mathbb{E}\left[\left|\left\{\mu(S_i,1,X_i,\mathrm{E}) - \mu(S_i,0,X_i,\mathrm{E})\right\}\right| \cdot \left|\frac{\rho(S_i,X_i)\cdot(1-\rho(S_i,X_i))}{\rho(X_i)\cdot(1-\rho(X_i))}\right| \middle| P_i = \mathrm{E}\right]$$

$$\leq c \cdot \mathbb{E}\left[\frac{\rho(S_i,X_i)\cdot(1-\rho(S_i,X_i))}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right]$$

The upper bound can be achieved by setting $\mu(s,0,x,\mathrm{E}) = \mu(s,x,\mathrm{O}) - c\cdot\rho(s,x)$ and $\mu(s,1,x,\mathrm{E}) = \mu(s,0,x,\mathrm{E}) + c$, These distributions are admissible under our assumptions, and hence sharpness is obtained. We can likewise obtain the lower bound.

□

Proof of Lemma 2 We show that the derived bounds are sharp by demonstrating that there exist data distributions consistent without assumptions that achieve these bounds. (i) In the absence of Comparability the data imply no restrictions on the values for $\mu(s,x,\mathrm{E})$, and so as long as there is some difference between $\rho(s,x)$ and $\rho(x)$ there is no bound on the bias.
(ii) If the outcomes are binary, the only restrictions implied on $\mu(s,x,\mathrm{E})$ are that all values lie inside $[0,1]$. The upper bound comes from imputing 1 for $\mu(s,x,\mathrm{E})$ if $\rho(s,x) > \rho(x)$ and 0 if $\rho(s,x) < \rho(x)$, a choice of distribution that is admissible. This directly implies the bounds on the bias.
(iii)

$$\text{comparability-bias} = \mathbb{E}\left[\left\{\mu(S_i,X_i,\mathrm{E}) - \mu(S_i,X_i,\mathrm{O})\right\}\cdot\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right].$$

Then

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\{\mu(S_i,X_i,\mathrm{E}) - \mu(S_i,X_i,\mathrm{O})\right\}\cdot\frac{\rho(S_i,X_i)-\rho(X_i)}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right]\right| \\
&\leq \mathbb{E}\left[\left|\left\{\mu(S_i,X_i,\mathrm{E}) - \mu(S_i,X_i,\mathrm{O})\right\}\right|\cdot\frac{|\rho(S_i,X_i)-\rho(X_i)|}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right] \\
&\leq c\cdot\mathbb{E}\left[\frac{|\rho(S_i,X_i)-\rho(X_i)|}{\rho(X_i)\cdot(1-\rho(X_i))}\middle| P_i = \mathrm{E}\right].
\end{aligned}$$

The upper bound can be attained by setting

$$\mu(s,x,\mathrm{E}) = \begin{cases} \mu(s,x,\mathrm{O}) + c & \text{if} \quad \rho(s,x) \geq \rho(x), \\ \mu(s,x,\mathrm{O}) - c & \text{otherwise,} \end{cases}$$

and similarly for the lower bound. □

C. Illustration of Bias Bounds Calculation

We will provide a simple illustration of how the theoretical bias bounds we calculated in Section 5.2 look like in practice. We focus on the employment outcome to illustrate the surrogacy bias and comparability bias bounds in the binary case (Case (ii)).

Table 9 shows the bounds on the treatment effects using the Influence Function Estimator under potential violations of Surrogacy and Comparability.[5] This demonstrates that in the binary outcome of employment, the sign can still be credibly inferred under the latter half even under surrogacy violation. The comparability bias seems to be non-negligible, part of our design of choosing Riverside (experimental data) due to its unique "jobs first" approach, in contrast to the "human capital" approach used in LA, San Diego, and Alameda counties (observational data). Further work must be done to ensure cases where comparability bias is minimal. We can similarly compute non-binary outcomes like Earnings, with some plausible range of user-specified parameter c (Case (iii) in Section 5.2).

Table 9: Bounds on the Influence Function Estimator for Employment Outcome

	Without Surrogacy		Without Comparability	
t	Lower Bound	Upper Bound	Lower Bound	Upper Bound
1	-0.80	0.36	-0.03	0.02
2	-0.64	0.18	-0.09	0.07
3	-0.46	0.12	-0.12	0.11
4	-0.36	0.09	-0.14	0.13
5	-0.29	0.07	-0.14	0.13
6	-0.25	0.06	-0.15	0.14
12	-0.13	0.03	-0.16	0.15
18	-0.09	0.02	-0.16	0.16
24	-0.07	0.01	-0.17	0.17
30	-0.05	0.01	-0.18	0.18
36	-0.04	0.01	-0.18	0.18

[5] If we are interested in conducting inference on the partial identification bounds, we can take the approach illustrated in, e.g., Imbens and Manski (2004); Molinari (2020).