NBER WORKING PAPER SERIES

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

Working Paper 26463
http://www.nber.org/papers/w26463

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2019

We are grateful for discussions with James Dailey, Lawrence Katz, Dylan Small, Scott Stern, and Liang Xu and for comments from numerous seminar participants. We thank Emanuel Schertz and James Stratton for outstanding research assistance. This research was funded through National Science Foundation Grant DMS-1502437, the Chan-Zuckerberg Initiative, the Bill & Melinda Gates Foundation, and the Overdeck Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w26463.ack

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

Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang

NBER Working Paper No. 26463
November 2019
JEL No. C01,J0

ABSTRACT

A common challenge in estimating the long-term impacts of treatments (e.g., job training programs) is that the outcomes of interest (e.g., lifetime earnings) are observed with a long delay. We address this problem by combining several short-term outcomes (e.g., short-run earnings) into a "surrogate index," the predicted value of the long-term outcome given the short-term outcomes. We show that the average treatment effect on the surrogate index equals the treatment effect on the long-term outcome under the assumption that the long-term outcome is independent of the treatment conditional on the surrogate index. We then characterize the bias that arises from violations of this assumption, deriving feasible bounds on the degree of bias and providing simple methods to validate the key assumption using additional outcomes. Finally, we develop efficient estimators for the surrogate index and show that even in settings where the long-term outcome is observed, using a surrogate index can increase precision. We apply our method to analyze the long-term impacts of a multi-site job training experiment in California. Using short-term employment rates as surrogates, one could have estimated the program's impacts on mean employment rates over a 9 year horizon within 1.5 years, with a 35% reduction in standard errors. Our empirical results suggest that the long-term impacts of programs on labor market outcomes can be predicted accurately by combining their short-term treatment effects into a surrogate index.

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Guido W. Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
Imbens@stanford.edu

Raj Chetty
Department of Economics
Harvard University
Littauer 321
Cambridge, MA 02138
and NBER
chetty@fas.harvard.edu

Hyunseung Kang
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue
Madison, WI 53706
hyunseung@stat.wisc.edu

Replication Code is available at https://opportunityinsights.org/data/

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¶

November 2019

Abstract

A common challenge in estimating the long-term impacts of treatments (e.g., job training programs) is that the outcomes of interest (e.g., lifetime earnings) are observed with a long delay. We address this problem by combining several short-term outcomes (e.g., short-run earnings) into a "surrogate index," the predicted value of the long-term outcome given the short-term outcomes. We show that the average treatment effect on the surrogate index equals the treatment effect on the long-term outcome under the assumption that the long-term outcome is independent of the treatment conditional on the surrogate index. We then characterize the bias that arises from violations of this assumption, deriving feasible bounds on the degree of bias and providing simple methods to validate the key assumption using additional outcomes. Finally, we develop efficient estimators for the surrogate index and show that even in settings where the long-term outcome is observed, using a surrogate index can increase precision. We apply our method to analyze the long-term impacts of a multi-site job training experiment in California. Using short-term employment rates as surrogates, one could have estimated the program's impacts on mean employment rates over a 9 year horizon within 1.5 years, with a 35% reduction in standard errors. Our empirical results suggest that the long-term impacts of programs on labor market outcomes can be predicted accurately by combining their short-term treatment effects into a surrogate index.

Keywords: Potential Outcomes, Causality, Surrogate Outcomes, Mediators, Propensity Score, Principal Stratification, Job Training, Earnings Impacts

*We are grateful for discussions with James Dailey, Lawrence Katz, Dylan Small, Scott Stern, and Liang Xu and for comments from numerous seminar participants. We thank Emanuel Schertz and James Stratton for outstanding research assistance. This research was funded through National Science Foundation Grant DMS-1502437, the Chan-Zuckerberg Initiative, the Bill & Melinda Gates Foundation, and the Overdeck 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

Estimating the long-term effects of treatments is of interest in many fields, from economics (e.g., the effects of early childhood interventions on lifetime earnings), to marketing (e.g., the effects of incentives on long-term purchasing behavior), to medicine (e.g., the effects of drugs on mortality rates). A central challenge in estimating such treatment effects is that long-term outcomes are often observed with a delay of many years or even decades, well beyond the time frame needed to make policy decisions.

One common approach to overcome this missing data problem is to analyze treatments effects on a short-term proxy variable, termed a "statistical surrogate" (Prentice 1989). The key assumption for a variable to be a valid surrogate is that the long-term outcome is independent of treatment conditional on the surrogate, which we term the "surrogacy assumption" (Begg and Leung 2000; Frangakis and Rubin 2002). For example, in studies of the effect of early childhood interventions (e.g., a class size reduction) on earnings, test scores serve as a statistical surrogate for earnings if earnings are independent of class size conditional on test scores, i.e., if any causal effect of class size on earnings is reflected in test scores. Under this assumption, the treatment effect on earnings can be identified from the causal effect of class size on test scores and the observational relationship between test scores and earnings. Intuitively, in a linear model, one can simply multiply the treatment effect of class size on test scores by the regression coefficient for the effect of test scores on earnings in an observational sample, a common practice in the education literature.

Unfortunately, the validity of the surrogacy assumption is difficult to verify and 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 appear to 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). Furthermore, any unmeasured confounding between the surrogate and long-term outcome would invalidate the surrogacy assumption, 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). Because of these issues, the use of intermediate outcomes as surrogates for long-term outcomes is often viewed with skepticism.

In this paper, we revisit the use of intermediate outcomes as surrogates from a different perspective. Rather than attempting to determine whether the surrogacy condition holds for any one intermediate outcome, we combine several intermediate outcomes into a "surrogate index". In a linear model, the surrogate index is simply the predicted value from a regression of the long-term outcome on the intermediate outcomes; more generally, the surrogate index is the conditional expectation of the long-term outcome given the intermediate outcomes (and any pre-treatment covariates). We show that the treatment effect on the long-term outcome can be identified as the treatment effect on the surrogate index under the assumption that the long-term outcome is independent of the treatment conditional on the full *set* of surrogates. This assumption weakens the standard surrogacy condition for a single variable: even if no one intermediate outcome satisfies the surrogacy assumption by itself, a set of intermediate outcomes together may together span the causal chain between the treatment and the long-term outcome. Of course, there is no guarantee that a large set of intermediate outcomes will together satisfy the surrogacy condition either; however, there is a greater likelihood that a set of outcomes will together satisfy the surrogacy assumption. We demonstrate the empirical relevance of this point using an application involving the estimation of long-term earnings impacts, in which any single short-term measure is not a valid surrogate, but a combination of short-term measures does turn out to satisfy surrogacy.

The approach of combining intermediate outcomes that we propose here is straightforward and intuitive, but virtually all existing empirical studies use a single predictive variable as a surrogate. Moreover, in the formal statistical literature on surrogacy, the relationship between the surrogate and the final outcome is simply postulated, and the statistical analysis focuses exclusively on the relationship between the treatment and the surrogate. The principal contributions of this paper are to (1) formalize how intermediate outcomes can be combined to efficiently estimate long-term impacts, making precise the assumptions under which such an approach yields unbiased estimates and formally incorporating estimation of the relationship between the surrogates and long-term outcome based on a secondary dataset; (2) develop feasible bounds on the degree of bias that can arise from violations of these assumptions as well as methods to validate the assumptions out-of-sample; (3) show how the use of a surrogate index can yield gains in efficiency even when long-term treatment effects can be estimated directly; and (4) present an empirical application showing that the use of multiple intermediate outcomes

rather than a single predictor can expedite the prediction of long-term treatment effects in the labor market by several years and increase precision significantly.

Formally, we study a setting with two samples, an "experimental sample" and an "observational sample." The experimental sample contains data about the treatment indicator and the surrogates but not the long-term outcome, which we term the *primary outcome*. The observational sample contains information about the surrogates and the primary outcome, but not the treatment indicator. Both samples may also contain pre-treatment variables. As an example, consider evaluating the effects of reductions in class size in elementary school on earnings in adulthood. Chetty et al. (2011) estimated the effect of class size and quality on earnings by linking data from the STAR experiment, which randomized class size in grades kindergarten to third grade in the 1980s, to information on earnings when children were in their mid-twenties. They extrapolated from these estimates to predict impacts on total lifetime earnings, assuming that the impacts on earnings in children's twenties would persist (in percentage terms) throughout their lives. Our goal is to develop methods to estimate lifetime earnings impacts without making such a strong assumption or waiting decades to observe lifetime earnings. In our framework, the experimental sample in this application would include data about class size (the treatment) and intermediate outcomes, such as test scores, college attendance rates, and earnings in early adulthood. The observational sample would be a large panel dataset that would include the same intermediate outcomes as well as earnings throughout adulthood, but not necessarily the treatment (class size).[1]

We study four questions. First, under what assumptions can the average treatment effect (ATE) on the long-term outcome be identified using surrogates? Second, what is the bias from violations of these assumptions and how can the assumptions be validated? Third, how can the average effect on the outcome be efficiently estimated using a vector of surrogates that collectively satisfy the surrogacy assumption? Fourth, if the primary outcome is also observed in the experimental sample, what information is gained by using surrogates?

To answer the first question, we introduce the *surrogate index*, defined as the expectation of the primary outcome conditional on the surrogates. We show that the difference in the mean

[1] As another example, an internet company may be interested in the causal effect of a change in the user experience on long-term engagement with a specific website, e.g., overall time spent on the website. Surrogates in that case could include detailed measures of medium-term engagement, including which of many webpages were visited and how long a user spent on each one.

surrogate index identifies the ATE on the primary (long-term) outcome when three assumptions hold: unconfoundedness, surrogacy, and comparability. Unconfoundedness is the familiar requirement that treatment is orthogonal to potential outcomes in the experimental sample (e.g., via random assignment of treatment). Surrogacy requires that the long-term outcome is independent of the treatment, conditional on the full set of surrogates. In the class size application discussed above, the surrogacy requirement is that the surrogates (e.g., test scores, college attendance, or earnings in early adulthood) together capture all of the effects of the class size intervention on the primary outcome (total lifetime earnings). Comparability requires that the conditional distribution of the primary outcome given the surrogates is the same in the observational and experimental samples. In the class size example, we require that the relationship between short-term earnings and lifetime earnings in the observational sample is the same as what one would have estimated in the experimental sample, were lifetime earnings observed there. Under these three assumptions, the ATE on the primary outcome is just identified and hence the assumptions jointly do not have any testable implications.

Next, we evaluate the degree of bias from the use of surrogates when the surrogacy condition fails.[2] In this case, our approach estimates the average causal effect on the conditional expectation of the primary outcome given the surrogate outcomes in the observational sample. We characterize the difference between this functional and the average treatment effect on the primary outcome itself. We then derive empirically estimable bounds on the degree of bias that are a function of the residual variances of the primary outcome and the treatment assignment indicator conditional on the surrogates, similar in spirit to recent expressions derived in standard treatment effect settings (e.g., Oster 2019, Andrews and Oster 2019). The bounds can be sharpened if one has information about the variance of the primary outcome conditional on treatment assignment, which cannot be directly estimated in sample but could be gauged from previous studies or other external evidence.

The analysis of the degree of bias shows why using many intermediate outcomes generally reduces the degree of bias. Intuitively, following the logic of directed acyclical graphs (DAGs, Pearl (1995, 2000)), the degree of bias is determined by the extent to which the intermediate

[2]We focus on the validity of the surrogacy assumption, as unconfoundedness is a widely-studied assumption in the literature on causal inference and comparability can typically be obtained by using an observational dataset that is well aligned with the experimental sample.

outcomes span the causal pathways from the treatment to the primary outcome. With a large and diverse set of intermediate outcomes, one is more likely to span all, or at least most of, these causal pathways. In the limiting case where the intermediate outcomes perfectly predict *either* the primary outcome *or* the treatment, the bias vanishes.

We also show how one can assess the validity of the surrogacy condition by treating one of the surrogates as a primary outcome and comparing the experimental estimate of the treatment effect on that outcome to the estimate based on the surrogate index. This out-of-sample validation approach is especially effective when the surrogates are the same as the long-term outcome but are measured earlier in time (e.g., short-term vs. long-term earnings). In such temporal settings, one can evaluate the surrogacy assumptions as time progresses by testing whether a surrogate index constructed based on early indicators tracks experimental outcomes well as time elapses, and then make longer-term predictions with greater confidence.

Third, we propose simple methods for estimating the average treatment effect under our identification assumptions. We show that the ATE on the primary outcome can also be estimated as the effect of the treatment on the surrogate index in the experimental sample. Thus, the surrogate index provides a simple way to collapse a high-dimensional vector of intermediate outcomes into a single index that can be used to estimate treatment effects. We also present alternative reweighting estimators that are based on a surrogate score, which have an interpretation analogous to propensity scores (Rosenbaum and Rubin 1983).

Fourth, we consider the case in which the researcher observes the primary outcome in the experimental sample itself so that one can directly identify the ATE on the primary outcome without making use of surrogates. Surrogates are helpful even in this setting: using the surrogate index, one can estimate the average effect of interest more precisely. Intuitively, the surrogate index reduces the residual variance of the outcome by eliminating variation that is orthogonal to the treatment. Building on the literature on semi-parametric estimation (e.g., Bickel et al. 1993), we characterize the efficiency gain from the use of the surrogate index. Surrogate indices increase precision the most in applications where the final outcome is a rare event (e.g., mortality) or where substantial noise is introduced after intermediate outcomes are measured. A somewhat surprising implication of these results is that it is typically preferable to use *fewer* surrogates to predict the long-term outcome in order to maximize precision (provided that the surrogacy assumption is satisfied).

In the final section of the paper, we apply our methods to re-analyze the impacts of a randomized trial of the GAIN job assistance program conducted by MDRC in the late 1980s, which provided job training and search services to unemployed individuals in California. We use experimental data from Hotz, Imbens and Klerman (2006) with nine years of post-program earnings from four urban counties: Alameda (Oakland), Los Angeles, Riverside, and San Diego. These sites implemented different types of programs, with Riverside focusing on a "jobs first" approach while other sites prioritized the development of human capital.

We begin by analyzing the Riverside program, which previous work has shown led to the largest impacts on employment and earnings over nine years. We investigate the amount of time required to estimate the long-term (nine-year) average effect of the program on employment using short-term employment rates as surrogates. We find that constructing a surrogate index based on employment rates in the first six quarters after the experiment yields a predicted impact on mean employment over nine years that is very similar to the actual estimated impact over nine years in Riverside. In contrast, the conventional approach of estimating treatment effects on mean employment rates requires waiting 6.25 years to obtain an estimate that falls within the 95% confidence interval of the mean impact over nine years. Using employment rates in any one quarter by itself as a surrogate also yields a highly biased estimate of the long-term impact. Intuitively, there are highly non-linear dynamics in employment rates over time, making it very useful to combine employment rates over multiple quarters to forecast long-term employment impacts accurately.

Next, we apply our methods to evaluate potential biases from violations of surrogacy. The point estimate of the treatment effect on mean employment rates over nine years can be bounded to be strictly positive within 5 quarters under plausible assumptions about the explanatory power of treatment assignment for the long-term outcome. Moreover, the estimates using a surrogate index based on the first six quarters of employment rates closely track actual observed mean employment impacts year-by-year, showing how the surrogacy condition could be validated as time elapses. Within 3-4 years after the experiment, researchers could have confidence in the validity of the surrogacy condition and thereby make extrapolations to longer-term impacts more confidently.

The surrogate index also yields a substantial increase in precision: even if one waits to observe outcomes over nine years, using employment rates in the first six quarters to create

a surrogate index yields a 35% reduction in the standard error of the estimate. Hence, it is actually preferable to discard the long-term employment data even if those data are available when the program is being evaluated. Intuitively, under the surrogacy assumption, variation in mean nine-year employment rates beyond that captured by the surrogate index is noise that is orthogonal to the treatment and reduces the precision of the estimate.

Finally, we show that the surrogate index we estimated using data from Riverside predicts the long-term treatment effects on employment and earnings in other sites accurately. The GAIN program implemented in Los Angeles had small effects on earnings over nine years; the programs in Alameda and San Diego had mid-sized effects; and the program in Riverside had the largest effect. Our six-quarter surrogate-index estimates closely match this pattern. This cross-site validation shows that surrogate indices estimated in a given application can provide reliable predictions in other settings as well.

In sum, a surrogate index based on six quarters of employment data allows us to estimate the long-term employment impacts of the GAIN program much more rapidly and precisely. This empirical finding, which we view as a central result of the paper, is useful not just in the context of the GAIN program itself, but more broadly for a large set of studies that seek to estimate long-term earnings and employment impacts, but only have the data to estimate shorter-term impacts. In particular, it suggests that short-term treatment effects of programs on labor market outcomes can provide accurate predictions of their long-term impacts if they are suitably combined. As noted above, existing studies typically make assumptions such as a constant percentage impact on earnings to predict lifetime earnings impacts (e.g., Krueger 1999; Chetty et al. 2011; Chetty, Hendren and Katz 2016; Hendren and Sprung-Keyser 2019). The surrogate index we construct provides a more disciplined and precise approach to constructing such forecasts.[3]

We recognize that the credibility of the surrogacy assumption may be questioned in any given application, especially when viewed in isolation. We therefore view the best path forward as building a "library" of surrogate indices in which researchers systematically catalog the smallest

[3]More ambitiously, our method could be applied to forecast earnings impacts of childhood interventions purely from pre-labor-market outcomes such as test scores and other data in childhood, further expediting the detection of long-term impacts. This would require finding a set of pre-labor-market indicators that satisfy the surrogacy condition. Our empirical results suggest that labor market outcomes in a short interval can serve as valid surrogates for long-term labor market trajectories; whether one can construct valid surrogates for earnings purely using pre-labor-market variables is an open question for future work.

set of surrogates that successfully match long-term outcomes of interest across several studies (e.g., earnings, mortality, educational attainment). If one establishes, for instance, that six quarters of employment 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.

Related Literature. In addition to the literature on surrogacy (Prentice 1989; Fleming and DeMets 1996; Begg and Leung 2000; Gilbert and Hudgens 2008; Weir and Walley 2006; Xu and Zeger 2001; D'Agostino, Campbell and Greenhouse 2006; Qu and Case 2006; Alonso et al. 2006; Lauritzen 2004; Day and Duffy 1996), this study also builds on and relates to the literatures on mediation and missing data. The mediation literature (Baron and Kenny 1986; van der Laan and Petersen 2004; Imai, Keele and Tingley 2010; VanderWeele 2015; Tchetgen Tchetgen and Shpitser 2014; Zheng and van der Laan 2012) 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 is to determine the relative magnitudes of the direct and indirect effects. In our surrogacy analysis, the mediator plays the role of the surrogate, and we focus on the case in which the direct effect is absent. The methods proposed in the mediation literature for estimating the indirect effect are not directly applicable to our two-sample case because we do not observe the outcome and the treatment for the same units, but they can be adapted to the one-sample setting as we discuss below.

This paper is related to the missing data literature (Little and Rubin 2014) in the sense that it addresses a situation in which the key variable of interest – the primary outcome – is missing in the experimental sample. Our analysis can be viewed as a special case of studies on combining data sets, e.g., Ridder and Moffitt (2007); Chen et al. (2008). In particular Rässler (2012, 2004) refer to this setting with one variable missing in one part of the 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 W_i to be a general random variable, rather than

a binary indicator as in our setup. We present a more formal discussion of how our assumptions and results relate to those in the mediation and missing data literatures in Appendix A.

The paper is organized as follows. The next section sets up the problem. Section 3 defines the surrogate index and establishes the conditions under which it yields unbiased estimates of treatment effects on the primary outcome. Section 4 presents formulas for bias when the surrogacy assumption fails, derives bounds on the degree of bias, and discusses how the surrogacy condition can be validated. Section 5 discusses estimation, while Section 6 presents results on the efficiency gain from using the surrogate index. Section 7 presents the empirical application. Section 8 concludes.

2 Setup

Consider a setting with two samples: an Experimental (E) sample and an Observational (O) sample. The experimental and observational sample contain observations on $N_{\rm E}$ and $N_{\rm O}$ units, 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 for the group to which unit i belongs.

For the $N_{\rm E}$ individuals in the experimental group, there is a single binary treatment of interest, $W_i \in \{0, 1\}$, and a primary outcome, denoted by Y_i. This outcome is not observed for individuals in the experimental sample. However, we do measure intermediate outcomes, which we refer to as surrogates (to be defined precisely in Section 3.2), denoted by S_i for each individual. 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 individual. These variables are 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.

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

Overall, the units are characterized by the values of the sextuple $(Y_i(0), Y_i(1), S_i(0), S_i(1), X_i, W_i)$. We do not observe the full sextuple for any units. Rather, for units in the experimental sample

we observe only the triple (X_i, W_i, S_i) with support $\mathbb{X}$, $\mathbb{W} = \{0, 1\}$, and $\mathbb{S}$ respectively. In the observational sample, we do not observe to which treatment the N_{O} individuals were assigned. We observe the triple (X_i, S_i, Y_i), with support $\mathbb{X}$, $\mathbb{S}$, and $\mathbb{Y}$ respectively. We summarize this data setup and our notation in Table 1.

We are interested in the average effect of the treatment on the primary outcome in the population from which the experimental sample is drawn:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|P_i = \text{E}],$$

or similar estimands, such as the average primary outcome for the treated units, or for some other subpopulation. For ease of exposition, we focus on estimating the ATE τ here. The fundamental problem in estimating τ in the experimental group is that the outcome Y_i is missing for all units in the experimental sample. To address this missing data problem, we exploit the observational sample and its link to the experimental sample through the presence of the surrogate outcomes S_i. Note that the surrogates, like the pre-treatment variables, are not of intrinsic interest. The average causal effect of the treatment on the surrogates, $\tau_S = \mathbb{E}[S_i(1) - S_i(0)|P_i = \text{E}]$, is of interest only insofar as it aids in estimation of τ.

3 Identification

In this section, we discuss three assumptions that together allow us to combine the observational and experimental samples and estimate the causal effect of the treatment on the primary outcome using a set of intermediate outcomes. The first assumption is unconfoundedness or ignorability, common in the program evaluation literature, which ensures that adjusting for pre-treatment variables leads to valid causal effects. The second assumption is the surrogacy condition, which we define more precisely below, and is the key condition that allows to use the surrogate variables to proxy for the primary outcome. The third assumption is comparability, which ensures that we can learn about relationships in the experimental sample from the observational sample. After stating these three assumptions, we present our main identification result, showing how the ATE on the primary outcome can be identified by combining intermediate outcomes under these assumptions.

3.1 Unconfoundedness

For the individuals in the experimental group, define the propensity score as the conditional probability of receiving the treatment: $e(x) = \text{pr}(W_i = 1 | X_i = x, P_i = \text{E})$. 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):

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

This assumption implies that in the experimental group, we could estimate the average causal effect of the treatment on the outcome Y_i by adjusting for pre-treatment variables, if the Y_i were measured.

3.2 Surrogacy and the Surrogate Index

Because the primary outcome is not measured in the experimental group, we exploit surrogates to identify the treatment effect of W on Y. The defining property of these surrogates S_i is the following condition:

Assumption 2. (Surrogacy)

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

Intuitively, the surrogacy condition requires that the surrogates fully capture the causal link between the treatment and the primary outcome. Figure 1 illustrates the content of this assumption using directed acyclical graphs to represent the causal chain from the treatment to the surrogate to the long-term outcome, as in Pearl (1995).[4] Panel A shows a DAG where the surrogacy assumption is satisfied by a single intermediate outcome S that lies on the causal chain between W and Y. Panel B shows an example where Assumption 2 is violated because there is a direct effect of the treatment on the outcome that does not pass through the surrogate.

[4] In Appendix A, we present an analogous discussion of how our assumptions relate to assumptions made in the mediation and missing data literatures.

Panel C shows our approach to addressing this problem: introducing multiple intermediate outcomes that together span the causal chain from W to Y. In this example, the three intermediate outcomes together span the causal chain from W to Y and hence can be combined to construct a surrogate index that captures long-term treatment effects. Importantly, one does not necessarily have to observe every intermediate outcome that lies on the causal chain between W and Y. For example, if a treatment (e.g., smaller class sizes) affects earnings by increasing both math and science aptitude, math scores by themselves could serve as a valid surrogate if math and science scores are perfectly correlated. The key requirement is that the set of intermediate outcomes together span the set of causal pathways, either because they themselves are the causal factors or because they are correlated with the causal factors.

It is instructive to compare the surrogacy assumption to the exclusion restriction assumption familiar to economists in instrumental variables settings. Figure 1d shows a DAG representation of the standard instrumental variables (IV) model, where there is an unobserved confounder between W and Y. In the standard IV approach, this confound is addressed by introducing an instrument Z that affects W but does not affect Y directly (the exclusion restriction). In the surrogacy case, we are interested in the effect of W on Y, where we assume there is no confounder between W and Y (or, equivalently, we find an instrument for W that eliminates such confounds). This is analogous to the (reduced-form) effect of Z on Y in the IV case. The reduced-form effect can be estimated directly in the IV case because Z and Y are both observed in the same dataset. The problem we address here is how to estimate the effect of Z (or W, assuming unconfoundedness) on Y when they are not observed in the same dataset. The analog to the exclusion restriction here is that there is no direct effect of W on Y that does not run through S.

We exploit the availability of multiple intermediate outcomes by defining two concepts: 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 in the observational sample:*

$$h_{\mathrm{O}}(s,x) = \mathbb{E}\left[Y_i \middle| S_i = s, X_i = x, P_i = \mathrm{O}\right].$$

The surrogate index $h_{\mathrm{O}}(s,x)$ is estimable because we observe the triple (Y_i, S_i, X_i) in the

observational sample.[5] In a linear model, the surrogate index is simply a linear combination of the individual intermediate outcomes – the predicted value from a regression of the primary outcome on the intermediate outcomes.

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:*

$$r(s,x) = \text{pr}(W_i = 1 | S_i = s, X_i = x, P_i = E).$$

Like the propensity score, the surrogate score facilitates statistical procedures that adjust only for scalar differences in other variables, irrespective of the dimension of the statistical surrogates.[6]

Proposition 1. (Surrogate Score) *Suppose Assumption 2 holds. Then:*

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

All proofs are given in Appendix B.

3.3 Comparability

Surrogacy and unconfoundedness by themselves are not sufficient for consistent estimation of τ by itself because they do not place restrictions on how the relationship between Y and S in the observational sample compares to that in the experimental sample. The final 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. Formally,

Assumption 3. (Comparability of Samples)

$$Y_i \,\Big|\, S_i, X_i, P_i = \text{O} \;\sim\; Y_i \,\Big|\, S_i, X_i, P_i = \text{E}.$$

[5] The conditional means we define in the surrogate index are related to what Hansen (2008) calls the prognostic score, although in the setting Hansen considers there is no surrogate variable, and the conditional expectation is only a function of the pre-treatment variables.

[6] In contrast to the familiar definition of the propensity score, we write the probability of "having received the treatment" rather than "receiving the treatment" because the surrogate score is conditional on a post-treatment outcome, whereas the propensity score conditions solely on pre-treatment variables.

We can state this assumption equivalently as:

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

To understand the role of the comparability assumption, note that there are two conditional expectations that are closely related to the conditional expectation in the definition of the surrogate index above, but which we cannot directly estimate because we do not observe Y in the experimental sample. The first is the conditional expectation corresponding to the definition of the surrogate index above within the experimental sample:

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

The second is the conditional expectation of the potential outcomes given pre-treatment variables and the surrogates:

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

These conditional expectations are all equivalent under comparability and surrogacy, allowing us to take the relationship between Y and S estimated in the observational sample and apply it in the experimental sample. In effect, 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 2 holds. Then:*

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

(ii) Suppose Assumption 3 holds. Then:

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

(iii) Suppose Assumptions 2 and 3 hold. Then:

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

Part (*iii*) of Proposition 2 relates the conditional expectation of interest, $\mu_{\rm E}(s,x,w)$, to a conditional expectation that is directly estimable, $h_{\rm O}(s,x)$.

Finally, we define weights that make the observational and experimental samples comparable. Let $q = N_E/(N_E + N_O)$ denote the sampling weight of being in the experimental sample and $(1-q)$ be the sampling weight of being in the observational sample. Define the propensity to be in the experimental sample $P_i = E$ as follows:

Definition 3. (Sampling Score)

$$t(s,x) = \text{pr}(\, P_i = E |\, S_i = s, X_i = x) = \frac{\text{pr}(S_i = s, X_i = x | P_i = \text{E}) q}{\text{pr}(S_i = s, X_i = x | P_i = \text{E}) q + \text{pr}(S_i = s, X_i = x | P_i = \text{O})(1-q)}.$$

We also make the assumption:

Assumption 4. Overlap in Sampling Score

$$t(s,x) < 1 \quad \text{for all } s \in \mathbb{S} \quad \text{and } x \in \mathbb{X}.$$

3.4 Identification

We now present our central identification result. We present three different representations of the average treatment effect that lead to three estimation strategies. The motivation for developing the different representations is that estimators corresponding to those different representations can have different properties in finite samples. The first representation requires estimation of the surrogate index, but not the surrogate score. The second representation instead requires estimation of the surrogate score, but not the surrogate index. The third representation requires estimation of both.

We define the following three objects, all functionals of distributions that are directly estimable from the data, starting with a surrogate index representation:

$$\tau^{\rm E} = \mathbb{E}\left[h_{\rm O}(S_i, X_i) \cdot \frac{W_i}{e(X_i)} - h_{\rm O}(S_i, X_i) \cdot \frac{1-W_i}{1-e(X_i)} \middle| P_i = \text{E} \right], \tag{3.2}$$

then a surrogate score representation,

$$\tau^{\rm O} = \mathbb{E}\left[Y_i \cdot \frac{r(S_i, X_i) \cdot t(S_i, X_i) \cdot (1-q)}{e(X_i) \cdot (1 - t(S_i, X_i)) \cdot q} \right. \tag{3.3}$$

$$-Y_i \cdot \frac{(1-r(S_i,X_i))\cdot t(S_i,X_i)\cdot(1-q)}{(1-e(X_i))\cdot(1-t(S_i,X_i))\cdot q}\,\Bigg|\, P_i = \mathrm{O}\Bigg],$$

and finally an influence function representation:

$$\tau^{\mathrm{O,E}} = \mathbb{E}\left[\psi(P_i, Y_i, S_i, W_i, X_i)\right], \tag{3.4}$$

where

$$\psi(p,y,s,w,x) = \frac{\mathbf{1}_{p=\mathrm{E}}}{q}\left(\frac{h_{\mathrm{O}}(s,x)w}{e(x)} - \frac{h_{\mathrm{O}}(s,x)(1-w)}{1-e(x)}\right) \tag{3.5}$$
$$+\frac{\mathbf{1}_{p=\mathrm{O}}}{1-q}\left(\frac{t(s,x)}{1-t(s,x)}\frac{1-q}{q}\right)\frac{\left(y-h_{\mathrm{O}}(s,x)\right)\left(r(s,x)-e(x)\right)}{e(x)(1-e(x))}.$$

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

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

The first representation, τ^{E}, shows how τ can be written as the expected value of the propensity-score-adjusted difference between treated and controls of the surrogate index in the experimental sample. This will lead to an estimation strategy in which the missing Y_i in the experimental sample are imputed by $\hat{h}(S_i, X_i)$ estimated on the observational sample. The second representation, τ^{O}, shows how τ can be written as the expected value of the difference in two weighted averages of the outcome in the observational sample, with the weights a function of the surrogate score estimated on the experimental sample and the sampling score. This will lead to an estimation strategy in which the Y_i in the observational sample are weighted proportional to the estimated surrogate score to estimate $\mathbb{E}[Y_i(1)|P_i = \mathrm{E}]$, and weighted proportional to one minus the estimated surrogate score to estimate $\mathbb{E}[Y_i(0)|P_i = \mathrm{E}]$. The third representation uses the score function representation, requiring estimation of both the surrogate score and the surrogate index.[7]

[7] In the standard unconfoundedness setting with observations on (W_i, Y_i, X_i), we can estimate the average treatment effect by estimating the propensity score, the conditional expectations of the potential outcomes, and the influence function. In that setting, the influence function approach has the attractive property that it is doubly robust in the sense that either the propensity score, or the conditional potential outcome expectations, can be completely misspecified without compromising consistency (Newey (1994); Chernozhukov et al. (2016)). Unfortunately, that is not the case here. If we use an estimator based on the influence function it will be inconsistent if either the surrogate score or the surrogate index is inconsistently estimated.

Under smoothness assumptions, we can derive the semi-parametric efficiency bound for τ (*e.g.*, Bickel et al. 1993; Newey 1990). Because the model is just identified (the model has no testable implications), it follows that the semi-parametric efficiency bound is the square of the influence function $\psi(\cdot) - \tau$:

$$\begin{aligned}\mathbb{V}_s &= \mathbb{E}\left[\left(\psi(P_i, Y_i, X_i, S_i, W_i) - \tau\right)^2\right] \\ &= \mathbb{E}\left[\frac{\sigma^2(S_i)}{1 - t(S_i, X_i)} \cdot \left(\frac{r(S_i, X_i)}{e(X_i)^2} + \frac{1 - r(S_i, X_i)}{(1 - e(X_i))^2} - 2 \cdot \frac{r(S_i, X_i) \cdot (1 - r(S_i, X_i))}{e(X_i)^2 \cdot (1 - e(X_i))^2}\right)\right. \\ &\qquad \left. + \frac{1}{t(S_i, X_i)} \cdot \left\{\frac{r(S_i, X_i)}{e(X_i)} \cdot \left(\mu(S_i, X_i) - \mu_1\right)^2 + \frac{1 - r(S_i, X_i)}{1 - e(X_i)} \cdot \left(\mu(S_i, X_i) - \mu_0\right)^2\right\}\right].\end{aligned}$$

Again because of the just-identified nature of this model, the results in Newey (1994) also imply that nonparametric estimators of the surrogacy score, the surrogacy index, and the propensity score can be used to obtain efficient estimators for τ.

4 The Surrogacy Assumption: Bias, Bounds, and Validation

In this section, we examine the biases that arise from violations of the surrogacy assumption. We first characterize the bias and then derive estimable bounds on the degree of the bias that can arise from such violations. Finally, we present a simple approach for validating the surrogacy assumption out-of-sample using a "hold out" intermediate outcome.

4.1 Bias from Violations of Surrogacy

We begin by characterizing the probability limit of estimators based on the characterizations of the estimand 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.

Theorem 2. (*i*) *Suppose Assumptions 1 (unconfoundedness) and 4 (overlap) hold, but Assumptions 2 (surrogacy) and 3 (comparability) do not necessarily hold. Then*

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

(ii) Suppose Assumptions 1 (unconfoundedness), 2 (surrogacy) and 4 (overlap) hold, but Assumption 3 (comparability) does not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\begin{aligned}
&\tau - \mathbb{E}\left[h_{\mathrm{O}}(S_i(1), X_i) - h_{\mathrm{O}}(S_i(0), X_i) \middle| P_i = \mathrm{E}\right] \\
&\quad = \mathbb{E}\left[\left\{h_{\mathrm{E}}(S_i, X_i) - h_{\mathrm{O}}(S_i, X_i)\right\} \cdot \frac{r(S_i, X_i) - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| P_i = \mathrm{E}\right].
\end{aligned}$$

(iii). Suppose Assumptions 1 (unconfoundedness), 3 (comparability) and 4 (overlap) hold, but Assumption 2 (surrogacy) does not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\begin{aligned}
&\tau - \mathbb{E}\left[h_{\mathrm{O}}(S_i(1), X_i) - h_{\mathrm{O}}(S_i(0), X_i) \middle| P_i = \mathrm{E}\right] \\
&\quad = \mathbb{E}\left[\left\{\mu_{\mathrm{E}}(S_i, X_i, 1) - \mu_{\mathrm{E}}(S_i, X_i, 0)\right\} \cdot \frac{r(S_i, X_i) \cdot (1 - r(S_i, X_i))}{e(X_i) \cdot (1 - e(X_i))} \middle| P_i = \mathrm{E}\right]
\end{aligned}$$

(iv). Suppose Assumptions 1 (unconfoundedness) and 4 (overlap) hold, but Assumptions 2 (surrogacy) and 3 (comparability) do not necessarily hold. Then the difference between the average causal effect and the estimand is

$$\begin{aligned}
&\tau - \mathbb{E}\left[h_{\mathrm{O}}(S_i(1), X_i) - h_{\mathrm{O}}(S_i(0), X_i) \middle| P_i = \mathrm{E}\right] \\
&\quad = \mathbb{E}\left[\left\{\mu_{\mathrm{E}}(S_i, X_i, 1) - \mu_{\mathrm{E}}(S_i, X_i, 0)\right\} \cdot \frac{r(S_i, X_i) \cdot (1 - r(S_i, X_i))}{e(X_i) \cdot (1 - e(X_i))} \middle| P_i = \mathrm{E}\right] \\
&\qquad + \mathbb{E}\left[\left\{h_{\mathrm{E}}(S_i, X_i) - h_{\mathrm{O}}(S_i, X_i)\right\} \cdot \frac{r(S_i, X_i) - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| P_i = \mathrm{E}\right].
\end{aligned}$$

Theorem 2(i) shows that even without surrogacy and comparability, we estimate a valid causal effect as long as unconfoundedness holds. The treatment effect we estimate is the average effect of the treatment on the surrogate index – an aggregate of intermediate outcomes – rather than the average effect on the primary outcome. This result also shows that the interpretation remains the same whether the surrogate estimate is constructed using the surrogate score approach or the surrogate index approach. The second, third, and fourth results show how a lack of comparability or surrogacy affect the difference between what is being estimated and the average treatment effect on the primary outcome.

The bias from violations of surrogacy (part (*iii*) of the theorem) – which we view as the assumption whose validity is most likely to be questioned in typical applications – consists of two factors. The first factor is small if the surrogates and pre-treatment variables explain a large share of the variation in Y_i and therefore $\mu_{\mathrm{E}}(s,x,1)$ and $\mu_{\mathrm{E}}(s,x,0)$ are close. In terms of the DAG shown in Figure 1c, the surrogates block most of the paths between W and Y, and hence the treatment effect on an index based on S must match that on Y.

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 $\mathbb{E}[r(S_i,X_i)\cdot(1-r(S_i,X_i))]$ is close to zero. This result may be more surprising: even if the surrogates do not capture much of the variation in the primary outcome, as long as the treatment and the surrogate are highly correlated, the bias from violations of surrogacy must be modest. Intuitively, if S captures all the variation in W, then the relationship between S and Y in the observational sample must coincide with the relationship between the W and Y. Since comparability guarantees that the relationship between S and Y is the same in both samples, it follows that we identify the relationship between W and Y in the experimental sample – the treatment effect of interest.

4.2 Bounds on the Magnitude of the Bias

We now show how one can derive empirically estimable bounds on the magnitude of the potential bias when surrogacy fails using the expressions in the previous subsection.

For simplicity, we consider the case where comparability holds with a constant sampling score, but surrogacy is violated. Assume that treatment assignment in the experimental sample is random and that there are no covariates X_i. Furthermore, suppose that the conditional expectation of the outcome give the surrogates and the treatment is linear and homoskedastic:

$$\mu_{\mathrm{E}}(S_i,W_i)\equiv\mathbb{E}\left[Y_i|S_i,W_i,p_i=\mathrm{E}\right]=\alpha+\gamma^\top S_i+\beta W_i,\qquad\text{and }\ \mathbb{V}\left(Y_i|S_i,W_i,p_i=\mathrm{E}\right)=\sigma^2_{Y|S,W}.$$

Here, β captures the direct effect of the treatment on the outcome that is not captured by the surrogates (as in Figure 1b).

Define the explanatory power of the surrogates for predicting the outcome as

$$R^2_{Y|S}=1-\frac{\sigma^2_{Y|S,W}+\beta^2\mathbb{E}[\mathbb{V}(W_i|S_i)]}{\mathbb{V}(Y_i)}=1-\frac{\sigma^2_{Y|S,W}+\beta^2\mathbb{E}[r(S_i)(1-r(S_i)]}{\mathbb{V}(Y_i)},$$

and similarly for predicting the treatment:

$$R^2_{W|S} = 1 - \frac{\mathbb{E}[r(S_i)(1-r(S_i)]}{\mathbb{E}[r(S_i)](1-\mathbb{E}[r(S_i)])} = 1 - \frac{\mathbb{E}[\mathbb{V}(W_i|S_i)]}{\mathbb{V}(W_i)},$$

where $\mathbb{V}(W_i|S_i)$ and $\mathbb{V}(W_i)$ are the conditional and marginal variances of W_i respectively, and $\mathbb{E}[\mathbb{V}(W_i|S_i)]$ is the expectation of the conditional variance which by definition is less than or equal to the marginal variance.

The bias in the surrogate index estimator is:

$$\begin{aligned}\text{bias} &\equiv \mathbb{E}\left[h_{\mathrm{O}}(S_i(1), X_i) - h_{\mathrm{O}}(S_i(0), X_i)\right| P_i = \mathrm{E}] - \tau \\ &= -\beta \times \frac{\mathbb{E}[r(S_i)(1-r(S_i)]}{\mathbb{E}[r(S_i)](1-\mathbb{E}[r(S_i)])} = -\beta \times \frac{\mathbb{V}(W_i|S_i)}{\mathbb{V}(W_i)} = -\beta \times (1 - R^2_{W|S}).\end{aligned}$$

Paralleling the intuition discussed above, the degree of bias depends upon the magnitude of the direct effect β and the residual variance in treatment assignment conditional on the surrogates. If $\beta = 0$, surrogacy is satisfied and there is no bias. If treatment assignment does not vary conditional on S, there is no scope for W to affect Y conditional on S and hence the surrogate index estimator must be unbiased.

The parameter $R^2_{W|S}$ can be estimated directly in the data. Hence, in order to obtain an estimable bound on the degree of bias, we need to bound β. To do so, note that the conditional expectation of Y_i given the surrogates is

$$\mathbb{E}\left[Y_i|S_i, p_i = \mathrm{E}\right] = \alpha + \gamma^\top S_i + \beta r(S_i).$$

It follows that β^2 can be bounded above by the R^2 for the explanatory power of the surrogates in explaining the primary outcome:

$$\beta^2 = \left\{\mathbb{V}(Y_i)\left(1 - R^2_{Y|S}\right) - \sigma^2_{Y|S,W}\right\} \Big/ \mathbb{E}[r(S_i)(1-r(S_i)] \leq \mathbb{V}(Y_i)\frac{1 - R^2_{Y|S}}{\mathbb{E}[r(S_i)(1-r(S_i)]}.$$

Since the square of the bias is

$$\text{bias}^2 = \left(\beta \times \frac{\mathbb{E}[r(S_i)(1-r(S_i)]}{\mathbb{E}[r(S_i)](1-\mathbb{E}[r(S_i)])}\right)^2,$$

it follows that

$$|\text{bias}| \leq \left(\frac{\mathbb{V}(Y_i)}{\mathbb{V}(W_i)} \times \left(1 - R^2_{W|S}\right) \times \left(1 - R^2_{Y|S}\right)\right)^{1/2}. \tag{4.1}$$

All of the variances and residual variances in this expression can be estimated, yielding an implementable upper bound on the degree of bias from violations of surrogacy. Intuitively, the bound on the bias is proportional to one minus the R^2 from the regression of the treatment on the surrogates and one minus the R^2 from the regression of the outcome on the surrogates. If the surrogates explain a large share of variation of the treatment or in the outcome, then the violations of the surrogacy assumption matter less, following the intuition for Theorem 2(*iii*) discussed above.

Sharpening the Bounds. We can sharpen the bounds by bringing in information on the explanatory power of the treatment W_i for the primary outcome Y_i given S_i. First define the combined explanatory power of the treatment W_i and the surrogates S_i:

$$R^2_{Y|S,W} = 1 - \frac{\mathbb{E}[\mathbb{V}(Y_i|S_i, W_i)]}{\mathbb{E}[\mathbb{V}(Y_i)]}.$$

Define the incremental explanatory power of the treatment over the surrogates as:

$$\delta = \frac{R^2_{Y|S,W} - R^2_{Y|S}}{1 - R^2_{Y|S}},$$

so that $\delta \in [0, 1]$. Then the bound on bias can be written as:

$$|\text{bias}| \leq \left(\frac{\mathbb{V}(Y_i)}{\mathbb{V}(W_i)} \times \delta \times \left(1 - R^2_{W|S}\right) \times \left(1 - R^2_{Y|S}\right) \right)^{1/2}. \tag{4.2}$$

Intuitively, if the treatment explains very little of the variance in the primary outcome conditional on the surrogates (i.e., δ is small), then there is little scope for bias from a direct effect of W on Y.

Equation (4.2) is not empirically implementable because $R^2_{Y|S,W}$ cannot be directly estimated in our setting, as we do not see Y and W together in the experimental sample. To make progress, we make the assumption that the incremental explanatory power of the treatment for the primary outcomes given the surrogates is smaller than the unconditional explanatory power of the treatment:

$$\delta = \frac{R^2_{Y|S,W} - R^2_{Y|S}}{1 - R^2_{Y|S}} \leq R^2_{Y|W}.$$

We view this as a relatively weak requirement that will hold in most applications: it simply requires that the explanatory power of the treatment for the long-term outcome does not increase (in percentage terms) when one includes information on intermediate outcomes. Under this assumption,

$$|\text{bias}| \leq \left(\frac{\mathbb{V}(Y_i)}{\mathbb{V}(W_i)} \times R^2_{Y|W} \times \left(1 - R^2_{W|S}\right) \times \left(1 - R^2_{Y|S}\right) \right)^{1/2} = \tau \left(\left(1 - R^2_{W|S}\right) \times \left(1 - R^2_{Y|S}\right) \right)^{1/2}. \tag{4.3}$$

Of course, $R^2_{Y|W}$ – the extent to which the treatment explains the outcome of interest – is not empirically estimable in our setting either. However, one may be able to place bounds on $R^2_{Y|W}$ (or the treatment effect τ) from prior studies. For instance, if there have been several evaluations of similar programs in the past and in no case has $R^2_{Y|W}$ exceeded 10%, then one may be comfortable assuming $R^2_{Y|W} < 0.1$ to gauge the potential degree of bias from violations of surrogacy. Alternatively, one can evaluate how much larger than previously observed treatment effects the true treatment effect would have to be in order to generate a given amount of bias in the surrogate index estimate.

4.3 Validation of Surrogacy Assumption

Although the bounds derived above can be useful in assessing worst-case scenarios and testing null hypotheses, they will often be wide given that they place relatively little structure on the relationship between the surrogates and the long-term outcome. As an alternative approach to gaining confidence in the estimates obtained using surrogate indices, one can assess the validity of the surrogacy condition by "holding out" an intermediate outcome. Intuitively, by treating one of the surrogates as a primary outcome, one can assess whether the surrogate index performs well out-of-sample.

Formally, suppose we have a vector of intermediate outcomes S_i that we partition into two components: $S_i = (S_{i1}, S_{i2})$, where S_{i2} is a scalar. We treat S_{i2} as a pseudo-primary outcome and use S_{i1} as the vector of surrogates for that outcome. That is, we treat the experimental data on (W_i, S_{i1}) as the experimental sample and the observational data on (S_{i1}, S_{i2}) as the observational sample. We can then construct a surrogate index for S_{i2} using S_{i1} and compare

the actual experimental estimate of the average treatment effect of W_i on S_{i2} to the estimate based on the surrogate index. If these two estimates do not match, we would likely question the validity of the surrogacy assumption in predicting the actual long-term outcome of interest Y_i. If in contrast the two estimates are well aligned, we would have more confidence in the surrogacy condition required to identify treatment effects on Y_i, especially if we find that the surrogate index and experimental estimates match for many intermediate outcomes S_{i2} that are similar to Y_i. The assumption underlying this approach to validation is that if S_{i1} satisfies surrogacy for the long-term outcome Y, it would satisfy surrogacy for the pseudo-outcome S_{i2} (and vice versa). The logic is analogous to "placebo tests" using pre-determined or exogenous variables that are commonly used to evaluate unconfoundedness in standard treatment effect settings; although balance on pre-determined variables does not guarantee unconfoundedness, violations of balance would typically reduce one's confidence that unconfoundedness holds.

This validation approach is especially informative when the surrogates have a sequential nature, e.g., when the intermediate variables S_{it} measure the same outcome as the long-term outcome earlier in time (e.g., employment rates in earlier periods, as in our empirical application below). In this temporal setting, one can evaluate the surrogacy assumption as time progresses by testing whether a surrogate index constructed based on early indicators tracks experimental outcomes well as time elapses, and then make longer-term predictions with greater confidence.

When the long term outcome Y_i is itself observed in the experimental sample – i.e., if sufficient time elapses in the sequential example above so that Y_i is observed – one can directly test the surrogacy assumption (Assumption 2) by testing if the estimates based on the surrogate index match the standard ATE on the long-term outcome. Equivalently, one can test Assumption 2 by testing if the treatment predicts the long-term outcome conditional on the surrogates. Although this test is not useful for the present application itself (as one would be able to estimate long-term impacts directly), it can be used to validate a set of surrogates for use in future studies in similar settings.

5 Estimation

In this section, we first present a simple estimator for the surrogate index, our primary approach to combining intermediate outcomes. We then discuss alternative estimators based on reweight-

ing strategies that are first-order equivalent but can have different finite-sample properties.

5.1 Surrogate Index

Suppose we estimate the surrogate index as $\hat{h}_{\rm O}(s,x)$. 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 (3.2), is the following difference of the two averages over the experimental sample:

$$\hat{\tau}^{\rm E} = \frac{1}{\sum_{i=1}^{N_{\rm E}} W_i/\hat{e}(X_i)} \sum_{i=1}^{N_{\rm E}} \hat{h}_{\rm O}(S_i,X_i)\cdot\frac{W_i}{\hat{e}(X_i)} \tag{5.1}$$

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

We refer to this as the surrogate index estimator. Note that compared to the representation in the theorem, 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 $h_{\rm O}(s,x)$ was based on a linear specification for the regression of the primary outcome on the intermediate outcome, $h_{\rm O}(s,x) = \gamma_0 + \gamma_S's + \gamma_X'x$, this leads to

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

where $\hat{\tau}_S$ is an estimator for $\mathbb{E}\left[S_i(1)-S_i(0)\right]$.

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:

$$\hat{\tau}^{\rm E} = \hat{\gamma}_S'(\overline{S}_1 - \overline{S}_0) = \overline{h}_1 - \overline{h}_0, \tag{5.2}$$

where $\overline{h}_1 - \overline{h}_0$ is the difference in the average predicted values of the surrogate index $\hat{h}_{\rm O}(s,x)$ in the treated and control samples.[8] That is, the treatment effect on the long-term outcome can

[8]This expression is also familiar from the mediation literature (*e.g.*, Baron and Kenny 1986). However, we emphasize that in general, there may be interactions between the surrogates and pre-treatment variables.

be estimated as the treatment effect on the predicted value of the long-term outcome based on the surrogates.[9]

5.2 Alternative Estimators

We now present three alternative estimators that are based on the other representations for τ in the main theorem. These estimators rely on reweighting observations rather than constructing a surrogate index. Although these estimators are all asymptotically equivalent, their behavior in finite samples can vary, and the best estimator can depend on the setting; see, for instance, similar issues in the context of estimating average treatment effects under unconfoundedness (Imbens and Wooldridge (2009); Abadie and Cattaneo (2018)).

Surrogate Score Estimator. We now use the second representation for τ in the main theorem to derive an alternative estimator. Let $\hat{e}(x)$, $\hat{r}(s,x)$, and $\hat{t}(s,x)$, be estimators for $e(x)$, $r(s,x)$, and $t(s,x)$ respectively. These may be nonparametric estimators, or simply estimators based on generalized linear models. For example we could specify

$$e(x) = \frac{\exp(\beta_0 + \beta_X' x)}{1+\exp(\beta_0 + \beta_X' x)}, \qquad r(s,x) = \frac{\exp(\alpha_0 + \alpha_S' s + \alpha_X' x)}{1+\exp(\alpha_0 + \alpha_S' s + \alpha_X' x)},$$

and

$$t(s,x) = \frac{\exp(\delta_0 + \delta_S' s + \delta_X' x)}{1+\exp(\delta_0 + \delta_S' s + \delta_X' x)},$$

estimated by maximum likelihood or method of moments. Once we have estimates $\hat{e}(x)$, $\hat{r}(s,x)$ and $\hat{t}(s,x)$, we can plug them into the sample analogs of the expected values in the main theorem.

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

$$\hat{\tau}^{\mathrm{O}} = \frac{1}{\sum_{i=1}^{N_{\mathrm{O}}} \omega_{1,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_{\mathrm{O}}} Y_i \cdot \omega_{1,\hat{r},\hat{e},\hat{t}} - \frac{1}{\sum_{i=1}^{N_{\mathrm{O}}} \omega_{0,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_{\mathrm{O}}} Y_i \cdot \omega_{0,\hat{r},\hat{e},\hat{t}}, \tag{5.3}$$

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

$$\omega_{w,\hat{r},\hat{e},\hat{t}} = \frac{\hat{r}(S_i,X_i)^w \cdot (1-\hat{r}(S_i,X_i))^{1-w} \cdot \hat{t}(S_i,X_i) \cdot (1-q)}{\hat{e}(X_i)^w \cdot (1-\hat{e}(X_i))^{1-w} \cdot (1-\hat{t}(S_i,X_i)) \cdot q}.$$

[9] When the surrogacy condition fails, equation (5.2) still identifies the treatment effect on a weighted average of intermediate outcomes under unconfoundedness. In this case, the surrogate index serves as a disciplined way to choose weights when aggregating different intermediate outcomes, based on their predictive power for a long-term outcome of interest.

Efficient Score Estimator. We can also base estimation on the efficient score given in (3.5). Given estimators for the propensity score, the surrogate score, and the sampling score, denoted by $\hat{e}(\cdot)$, $\hat{r}(\cdot,\cdot)$, the surrogate index $\hat{h}(\cdot,\cdot)$, and $\hat{t}(\cdot,\cdot,\cdot)$, we can estimate the average treatment effect as

$$\hat{\tau}_{E,O} = \sum_{i=1}^{N} \frac{\mathbf{1}_{P_i=\mathrm{E}}}{\hat{q}} \left(\frac{\hat{h}_{\mathrm{O}}(S_i,X_i)W_i}{\hat{e}(X_i)} - \frac{\hat{h}_{\mathrm{O}}(S_i,X_i)(1-W_i)}{1-\hat{e}(X_i)} \right)$$

$$+ \frac{\mathbf{1}_{P_i=\mathrm{O}}}{1-\hat{q}} \left(\frac{\hat{t}(S_i,X_i)}{1-\hat{t}(S_i,X_i)} \frac{1-\hat{q}}{\hat{q}} \right) \frac{\left(Y_i - \hat{h}(S_i,X_i)\right)\left(\hat{r}(S_i,X_i) - \hat{e}(X_i)\right)}{\hat{e}(X_i)(1-\hat{e}(X_i))}.$$

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 must be first-order equivalent.

Double Matching Estimator. Finally, we consider a matching estimator. Although matching estimators are generally not efficient in settings with unconfoundedness (Rubin 2006; Abadie and Imbens 2006, 2016), they are intuitive and widely applied, and it is instructive to see how a matching strategy can be used here.

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 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 $i'(i)$ 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:

$$\widehat{Y_i(1) - Y_i(0)} = Y_{i'} - Y_{j'}. \tag{5.4}$$

The matching estimator for τ would then be the average value of (5.4) over the experimental sample.

5.3 Estimation with Many Potential Surrogates

When the number of pre-treatment variables or surrogates (and their interactions) is large, standard regularization methods such as LASSO (Tibshirani 1996; Belloni, Chernozhukov and Hansen 2014), ridge regression, tree or forest based methods (Breiman 2001; Wager and Athey 2018), or super learners (Van der Laan and Rose 2011) can be used to obtain better estimates of the relevant conditional expectations.

Similar to the familiar case of estimating average treatment effects under a selection on observables assumption, in which pre-treatment variables play two roles in achieving unconfoundedness, surrogates also play two roles, both of which should be taken into account in the variable selection procedure. Recall that in the standard case, pre-treatment variables are associated with the potential outcomes and with the treatment indicator. Biases from omitting pre-treatment variables arise from the combination of the association with the potential outcomes and the association with the treatment indicator. As a result, variable selection procedures need to take both into account (e.g., Belloni, Chernozhukov and Hansen 2014), and the most effective procedures involve doubly robust methods (e.g., Scharfstein, Rotnitzky and Robins 1999). In the current setting, the surrogates play two roles as well. Their association with the treatment indicator is captured through the surrogate score, while their association with the outcomes is captured in the surrogate index. Surrogate selection should therefore focus on selecting intermediate outcomes that are either strongly linked to the primary outcome or the treatment.[10]

[10]The latter requirement is closely related to the requirement that there are substantial effects of the treatment on the surrogates, but it is not the same. If two potential surrogates are highly correlated, there may be substantial treatment effects for both of them, but it may not be necessary to include both.

6 Efficiency Gains from Surrogate Indices

In this section, we show how surrogate indices increase the precision of treatment effect estimates in addition to expediting their estimation.

To quantify efficiency gains, we consider a single sample setting in which there is one (experimental) sample in which the long-term outcome is observed alongside treatment and the surrogates. Formally, all units in the population are characterized by the sextuple $(Y_i(0), Y_i(1), S_i(0), S_i(1), X_i, W_i)$. We observe the quadruple (S_i, X_i, W_i, Y_i), now including the realized outcome Y_i.

In this setting, it is well-known that an efficient estimator for the effect of a treatment W_i on Y_i is the difference between the sample mean of the treated outcomes and the sample mean of the control outcomes in the absence of covariates or further assumptions about the data generating process. Thus, it might seem that incorporating surrogate variables S_i in estimation (for example, by replacing Y_i with the surrogate index in estimation, as in τ^{E}) would reduce efficiency. However, the opposite is true under the surrogacy assumption. Intuitively, the surrogacy assumption brings additional information to bear on the problem – namely that any variation in Y conditional on S is orthogonal to W and hence is simply noise that increases the residual variance of the outcome and reduces precision. By identifying suitable surrogates – outcomes that capture the treatment effect of interest earlier in the causal chain shown in Figure 1c – one can strip out the residual variation arising from downstream factors that create noise. For example, if a program affects individuals' labor market trajectories entirely by changing their first job placement, then any subsequent changes in employment outcomes simply add noise. Hence, one can estimate the program's long-term effect on earnings more precisely by focusing solely on the portion of lifetime earnings that projects onto characteristics of an individual's first job.

To formalize this result, let $\sigma^2(s,x) = \mathbb{V}(Y_i|S_i = s, X_i = x)$, $\sigma_w^2(x) = \mathbb{V}(Y_i|X_i = x, W_i = w)$, and $\mu_w(x) = \mathbb{E}[Y_i|X_i = x, W_i = w]$.

Theorem 3. (*i*) *The efficiency bound without assuming surrogacy, but when surrogacy holds is*

$$\begin{aligned}\mathbb{V}_{\mathrm{ns}} &= \mathbb{E}\left[\frac{\sigma_1(X_i)^2}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \tau)^2\right] \\ &= \mathbb{E}\left[\sigma^2(S_i, X_i) \cdot \left(\frac{r(S_i, X_i)}{(e(X_i))^2} + \frac{1 - r(S_i, X_i)}{(1-e(X_i))^2}\right)\right.\end{aligned}$$

$$+\frac{r(S_i)}{(e(X_i))^2}\cdot\left(h_{\mathrm{E}}(S_i,X_i)-\mu_1(X_i)\right)^2+\frac{1-r(S_i,X_i)}{(1-e(X_i))^2}\cdot\left(h_{\mathrm{E}}(S_i,X_i)-\mu_0(X_i)\right)^2$$
$$+\left(\mu_1(X_i)-\mu_0(X_i)-\tau\right)^2\Bigg].$$

(ii) The efficiency bound assuming surrogacy is

$$\mathbb{V}_{\mathrm{s}}=\mathbb{E}\Bigg[\sigma^2(S_i,X_i)\cdot\left(\frac{r^2(S_i,X_i)}{(e(X_i))^2}+\frac{(1-r(S_i,X_i))^2}{(1-e(X_i))^2}\right)$$
$$+\frac{r(S_i)}{(e(X_i))^2}\cdot\left(h_{\mathrm{E}}(S_i,X_i)-\mu_1(X_i)\right)^2+\frac{1-r(S_i,X_i)}{(1-e(X_i))^2}\cdot\left(h_{\mathrm{E}}(S_i,X_i)-\mu_0(X_i)\right)^2$$
$$+\left(\mu_1(X_i)-\mu_0(X_i)-\tau\right)^2\Bigg].$$

The difference between the two bounds, $\mathbb{V}_{\mathrm{ns}}-\mathbb{V}_{\mathrm{s}}$, is the efficiency gain from exploiting surrogacy. The expressions differ in the first term, involving $\sigma^2(S_i,X_i)$. There is no gain if $S_i=W_i$ (the treatment can be perfectly inferred from the surrogates), or if $\sigma^2(s,x)=0$ (the final outcome can be inferred perfectly from the surrogates and pre-treatment variables). Intuitively, in these cases the surrogate index estimate is identical to the conventional experimental estimate as there is no residual variance in the outcome or treatment conditional on the surrogates.

To gain more intuition about the determinants of the efficiency gain, consider the case where $\sigma^2=\sigma^2(s,x)$ for all s (homoskedasticity) and there are no pre-treatment variables. In this case,

$$\mathbb{V}_{\mathrm{ns}}-\mathbb{V}_{\mathrm{s}}=\mathbb{E}\left[\frac{2\sigma^2}{p(1-p)}\cdot\left\{p(1-p)-\left(r(S_i)-p\right)^2\right\}\right],\tag{6.1}$$

where $p=\mathbb{E}[W_i]$. Again, there is no gain if $S_i=W_i$, in which case the surrogate score $r(S_i)\in\{0,1\}$. The efficiency gain is largest if $r(S_i)$ is constant (and thus equal to $\mathbb{E}[r(S_i)]=p$) and when the residual variance in the primary outcome Y given the surrogates σ^2 is large. Intuitively, the use of surrogates purges the most noise when the residual variances in W and Y given S are high, thereby yielding larger efficiency gains. Insofar as the unexplained variance is particularly large for rare outcomes, the use of surrogates is likely to yield particularly large precision gains when studying rare outcomes such as mortality (Athey and Stern 1998).

There is an interesting tradeoff between the settings in which surrogates yield large efficiency gains and settings in which violations of the surrogacy assumptions generate minimal bias.

Theorem 2 shows that the bias from violations of surrogacy is smallest when $r(S_i)$ is close to 0 or 1 or when σ^2 is small. Theorem 3 shows that the efficiency gains are largest under precisely the opposite conditions. Surrogates are thus likely to yield the largest gains in precision when the surrogacy assumption is known to be highly credible (e.g., based on out-of-sample validation as in Section 4.3), but there is considerable residual variance in treatment assignment and outcomes conditional on the surrogates.

The results in this section imply that when given a choice between many potential surrogates, it is optimal to use the *smallest* set of surrogates that satisfy the surrogacy assumption to maximize efficiency. In a temporal setting where surrogates are ordered sequentially over time, this implies that the objectives of rapid detection and maximizing precision are generally aligned: it is preferable to use outcomes over the shortest time period both to save time and maximize precision.

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 employment rates (Riccio et al. 1989; Friedlander and Robins 1995). GAIN was a job assistance program implemented in California in the late 1980s to help welfare recipients find work. 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 found; in contrast, other sites focused more heavily on developing human capital through training programs (Hotz, Imbens and Klerman 2006).

We begin by presenting a brief summary of the experimental design and describing how we construct a surrogate index estimator using short-term employment rates. We then illustrate each of 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. Finally, we assess whether the surrogate index accurately predicts the heterogeneity in long-term treatment effects across other sites.

7.1 The GAIN Program

The GAIN treatment was randomly assigned to welfare (Aid for Families with Dependent Children) recipients. This is a very low-income population. In Riverside, only 22% of the participants were employed and mean quarterly earnings were $453 in the quarter prior to random assignment. The treatment group consisted of $N_1 = 4405$ participants, which the control group consisted of $N_0 = 1040$ participants who were not eligible for the additional services. Let $N = N_1 + N_0 = 5445$ denote the total number of participants in the Riverside trial.

Hotz, Imbens and Klerman (2006) followed study participants for nine years after assignment of the treatment, measuring quarterly employment rates and earnings 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 2, which plots employment rates (Panel A) and quarterly wage earnings (Panel B) by quarter for individuals in the treatment and control groups.

In Riverside, the mean impact over the nine years after the experimental intervention was a 6.37 (s.e. = 1.1) percentage point (pp) increase in employment rates. Our question is whether this nine-year employment impact could have been estimated more quickly by using short-term employment impacts as surrogates. We use data drawn directly from Hotz, Imbens and Klerman (2006); see that paper for further details on data construction and Friedlander and Robins (1995) for further details on the GAIN program.

7.2 Surrogate Index Estimator

Let Y_{it} represent an indicator for whether individual i is employed in quarter t, for $t = 1, \ldots, 36$ after the experimental intervention. Let W_i represent a treatment indicator for individual i, where $W_i = 1$ denotes that individual i received the GAIN treatment and $W_i = 0$ denotes that he/she did not receive treatment. Let $\bar{Y}_{iT} = \sum_{i=1}^{T} Y_{it}/T$ denote the fraction of quarters for which individual i is employed from periods 1 to T.

Our goal is to estimate the treatment effect of the GAIN program on mean employment

rates over T quarters after the experiment. One could estimate this parameter as

$$\hat{\tau}_T = \frac{1}{TN_1}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}W_i - \frac{1}{TN_0}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}(1-W_i),$$

if one were willing to wait T quarters, as Hotz, Imbens and Klerman (2006) do using data for $T = 36$ quarters (nine years). Our goal is to expedite identification of this treatment effect by constructing surrogate indices using data on employment rates up to quarter $S < T$.

We construct a surrogate index for mean employment rates from quarters 1 to T for each individual based on observed employment indicators Y_{it} from quarters 1 to S. To do so, we estimate the following linear regression model using OLS, using only data from the treatment group:[11]

$$\bar{Y}_{iT} = \beta_0 + \sum_{t=1}^{S}\beta_t Y_{it} + \varepsilon_i. \tag{7.1}$$

The predicted value from this regression, which we denote by $\hat{Y}_{i,T,S}$, is our surrogate index for mean employment from quarters 1 to T based on the individual's employment record up to quarter S. We compute this surrogate index for each of the individuals in the experimental sample and define our estimate of the treatment effect based on the surrogate index as

$$\hat{\tau}_{T,S} = \frac{1}{N_1}\sum_{i=1}^{N}\hat{Y}_{i,T,S}W_i - \frac{1}{N_0}\sum_{i=1}^{N}\hat{Y}_{i,T,S}(1-W_i), \tag{7.2}$$

simply replacing the actual employment rate observed from periods 1 to T with the prediction based on the surrogate index, as in equation (5.2).

Because the data in this application come from a randomized experiment, the unconfoundedness assumption is satisfied in this setting. The comparability assumption is also satisfied because the treatment group (which we use as the "observational sample" to estimate the surrogate index) is a random sample of the full population. Hence, the key assumption in question

[11] We discard the control group when estimating the surrogate index to ensure that none of the variation used to estimate the surrogate index is itself experimentally induced, as would be the case in an application where one lacked experimental data for 9 years. In most applications, one would use historical data – e.g., observational data for the nine years preceding the experiment – to construct the surrogate index. We use the same treatment group data to evaluate the performance of the surrogate index vs. the actual experimental estimates, which could raise concerns about overfitting due to re-use of the same sample. We find that using a randomly split sample approach yields very similar estimates, and therefore report estimates based on this simple approach here.

is the surrogacy condition: how many quarters of data S on employment rates are adequate to capture mean treatment effects on employment rates over the full 9 year period – i.e., how much time and precision could we have gained in predicting long-term impacts by using a surrogate index?

7.3 Results for Riverside

Gains in Time. To calculate the potential gain in time that could have been achieved using a surrogate index, we compare the estimates obtained from our surrogate index estimator with the actual experimental estimates of the mean treatment effect in Riverside over 9 years, varying the surrogacy window S from 1 to 36. The estimates based on the surrogate index, plotted in the solid circles in Figure 3a, quickly converge to the nine-year mean impact. The point estimates from the surrogate index approach fall within the 95% confidence interval for the experimental estimate of the nine-year employment effect for $S \geq 3$ (three quarters of data). At $S = 6$, the point estimate of 6.73% nearly matches the actual experimental estimate of 6.37%. Using more than six quarters of data to construct the surrogate index yields estimates that remain relatively stable around 6%, suggesting that it is adequate to have 1.5 years of data to predict the nine year impact accurately.

By contrast, the point estimates from a naive approach of simply focusing on mean employment rates over the observed sample period – shown by open circles – fall within the 95% confidence interval for the experimental estimate only after the 25th quarter (6.25 years).[12] Moreover, we find that using employment rates in any one quarter by itself as a surrogate would generally yield a highly biased estimate of the long-term impact (Figure 3b). Because there are highly non-linear dynamics in employment rates over time – as shown in Figure 2 – combining employment rates over multiple quarters yields more accurate forecasts of long-term employment impacts. In particular, the surrogate index places much larger weight on employment in the most recent quarter relative to earlier quarters (Appendix Table 1).

Appendix Figure 2 shows that we find similar results when predicting quarterly earnings rather than employment rates: a surrogate index based on earnings over the first six quarters

[12]Including additional variables as predictors in the surrogate index – e.g., pre-treatment employment rates, racial demographics, education levels, etc. – has little impact on the estimates because these variables have very limited explanatory power for long-term employment rates above and beyond employment rates in a few quarters after random assignment.

after the experiment does well in forecasting nine-year mean impacts, while naive estimators based purely on short-term treatment effects perform poorly.

Validation. In Figure 4, we implement the out-of-sample validation approach discussed in Section 4.3. We examine how a surrogate index based on $S = 6$ quarters of post-experimental data performs in predicting mean employment impacts over varying horizons from $T = 6$ to $T = 36$ quarters. We find that the surrogate index estimates closely track the actual experimental estimates; in particular, the point estimates from the surrogate index approach always fall within the 95% confidence intervals for the experimental estimate, and the point estimates are generally very similar. Appendix Figure 3 shows similar results for earnings levels.[13]

Figure 4 shows how the surrogate index estimator can be validated out-of-sample before one makes longer-term predictions. The fact that the surrogate index estimates based on 6 quarters of data closely track experimental impacts 2, 3, and 4 years after random assignment serves to validate the key surrogacy assumption within a few years after the experiment. One can therefore make extrapolations to longer-term (e.g., 9-year) impacts with greater confidence a few years after random assignment.

Bounds. In Figure 5, we plot the bounds in (4.3), varying the number of quarters used to estimate the surrogate index as in Figure 3a. To select the key exogenous parameter $R^2_{Y|W}$ that governs the bounds, note that $R^2_{Y|W} = 0.66\%$ in our experimental sample.[14] We therefore consider values of $R^2_{Y|W} = 1\%$ and $R^2_{Y|W} = 5\%$, a more conservative assumption in which the explanatory power of the treatment for the long-term outcome is permitted to be five times as large as what is observed empirically.[15]

The bounds shrink as the number of surrogates used rises, collapsing to zero when $S = 36$ and the surrogate index fully explains the outcome of interest. When we assume $R^2_{Y|W} \leq 1\%$, the point estimate for the bounds excludes zero when $S \geq 5$, implying that one can rule out a zero

[13] We can also use the surrogate index approach to predict employment and earnings impacts in a given year, rather than the cumulative mean starting in the quarter after random assignment. We present results on estimated experimental impacts and surrogate index predictions by year in Appendix Figure 3. The surrogate index closely tracks the experimental estimates by year as well, although the surrogate index predictions deviate slightly more from the experimental estimates in later years.

[14] The conservative bounds that do not make any assumption about $R^2_{Y|W}$ in (4.1) are ten times as wide as the bounds shown in Figure 3a when $R^2_{Y|W} = 0.01$ and hence prove to be uninformative in this application.

[15] These choices are intended to parallel how one might choose $R^2_{Y|W}$ in actual applications: one could use the largest value observed in previously conducted studies or some multiple of that value (e.g., 5 times as large).

mean treatment effect over 9 years within 5 quarters after the experiment.[16] Under the more conservative $R^2_{Y|W} \leq 5\%$ assumption, the bounds exclude zero when $S \geq 18$. Hence, even if one is uncertain about the validity of the surrogacy conditions and the magnitude of the program's long-term impact, one can be confident that there is a positive impact on mean employment rates relatively quickly under plausible assumptions based on historical data. Intuitively, employment rates have sufficiently high serial correlation that there is not too much scope for long-term treatment effects to deviate from the predictions based on short-term outcomes even if the surrogacy condition is violated. Appendix Figure 4 shows similar results when predicting quarterly earnings rather than employment rates.

Gains in Precision. In addition to being available more rapidly, the surrogate index estimator is also much more precise, consistent with the results in Section 6. The standard error on the conventional experimental estimate of the treatment effect on average employment rates over 9 years is 1.1 percentage points, 54% larger than the standard error the corresponding estimate using the surrogate index estimator based on the first six quarters of employment (0.7 percentage points).[17] Equivalently, starting from the conventional experimental estimate as the reference, the surrogate index estimator reduces the standard error of the treatment effect by 35%. Likewise, we find that the standard error on the conventional experimental estimate of the treatment effect on average earnings over 9 years is 55% larger than an estimate based on a surrogate index constructed using earnings in the first six quarters after the experiment.

Using more than six quarters of data to predict long-term impacts yields strictly less precise estimates (Appendix Table 2). Hence, using more data to estimate the program's impacts is undesirable even if those data are available when the program is being evaluated. One intuition for why a surrogate index based on a few quarters of data increases precision is that the GAIN program largely operates by putting individuals on different trajectories in the first few quarters

[16] To simplify the figure, we ignore estimation error in the bounds. It is straightforward to construct confidence intervals for the bounds following Imbens and Manski (2004); for instance, a 95% coverage region can be constructed as the upper bound plus 1.645 times the standard error of the upper bound estimate and the lower bound minus 1.645 times the standard error. The standard errors for the point estimates for the bounds range from 0.3 to 1.0 percentage points as we vary S; as a result, the 95% coverage region is only slightly wider than the bounds plotted in Figure 5. At the lower bound of the 95% coverage region, we can rule out a zero effect starting in quarter 8 rather than quarter 5.

[17] To simplify computation, we construct confidence intervals simply based on the treatment effect estimate on the surrogate index, ignoring the estimation error in the surrogate index itself. Using a bootstrap to account for this source of error yields very similar estimates.

after the intervention. All subsequent fluctuations in employment and earnings are largely orthogonal to the initial treatment and hence simply add noise.

7.4 Results for Other Sites

We now turn to the data from other sites beyond Riverside.[18] Hotz, Imbens and Klerman (2006) report experimental estimates over a nine-year horizon in three other urban counties: Alameda (Oakland), Los Angeles, and San Diego. They show that the treatment effects of the GAIN program were very heterogeneous across sites, perhaps because they took different approaches to job training and served different types of populations.

Here, we ask whether one could have predicted this treatment effect heterogeneity on long-term outcomes more quickly through a surrogate index. Using the regression coefficients estimated in (7.1) in the Riverside sample, we construct predicted rates of mean employment over $T = 36$ quarters for individuals in each of the other sites using $S = 6$ quarters of data after the experiment.[19] We then estimate the treatment effect based on the surrogate index for each site separately as in (7.2).

Figure 6 plots treatment effect estimates based on a six-quarter surrogate index against the actual treatment effects estimated using nine years of data. Panel A shows results for employment rates, while Panel B shows results for quarterly earnings. The surrogate index estimates are closely aligned with the long-term experimental estimates across the sites. In particular the six-quarter surrogate index captures the finding that the GAIN treatment had much larger long-term effects on employment in Riverside than in other sites. The surrogate index also captures the fact that the earnings gains were largest in Riverside, smallest in Los Angeles, and in the middle in Alameda and San Diego.

Intuitively, the surrogate index performs well because sites with larger and more persistent short-term effects on employment and earnings tended to have larger long-term effects on employment and earnings. Because the trajectory of short-term outcomes is highly predictive of

[18] We analyze data from the other sites separately from Riverside to conduct an out-of-sample validation analysis. We analyzed the data from the other sites only after reporting the results above for Riverside, simply evaluating the performance of the six-quarter surrogate index already estimated in Riverside in the other sites.

[19] Since we use the coefficients estimated in the Riverside data to construct surrogate indices in the other sites, this analysis is a joint test of the surrogacy and comparability assumptions. In practice, we find that the surrogate indices one obtains are virtually identical regardless of which site's data are used, showing that these sites are quite comparable in terms of employment and earnings dynamics.

long-term outcomes in the observational sample in Riverside (Appendix Table 1), we obtain predicted treatment effects that closely mirror the actual long-term estimates. This cross-site validation demonstrates how surrogate indices estimated in a given setting can provide reliable predictions in other settings.

In summary, our analysis of the GAIN experiment shows that surrogate indices can yield substantial gains in time and precision. This result is of interest not just in the context of the GAIN evaluation itself, but a much broader set of studies that estimate treatment effects on employment and earnings using relatively short panels. As discussed in the introduction, many recent studies use estimates of short-term treatment effects on earnings to make predictions about lifetime earnings impacts, relying on strong assumptions – such as constant effects over time – to make such extrapolations. Our results suggest that these short-term treatment effects could be combined into a surrogate index that would yield more accurate predictions of long-term impacts using existing observational data.

8 Conclusion

This paper has proposed a simple method of 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 easily estimated using regression or other standard 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 expedites 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.

ONLINE APPENDICES

A. Connection to Mediation and Missing Data Literatures

In this appendix, we link the setup and assumptions in the current paper to the literatures on mediation and missing data.

Mediation. 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 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 $Y_i(w)$ can be defined as $Y_i(w) = Y_i(w, S_i(w))$, for $w \in \mathbb{W}$.)

In the setting considered in the mediation literature, we observe the quadruple (W_i, S_i, Y_i, X_i) for all units in the sample and so there is no distinction between the experimental sample and the observational sample. Hence, the comparability assumption is automatically satisfied (Frangakis and Rubin 2002; Mealli and Mattei 2012; Ding and Lu 2017).

The focus in 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 concepts are the average *total effect,*

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

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

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

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

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

with the 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 the ones made in Section 3. The first assumption in the mediation framework is a reformulation

of the unconfoundedness assumption, Assumption 1. It rules out the presence of unmeasured confounders between the treatment and the surrogate, and between the treatment and the outcome.

Assumption 5.

$$W_i \perp\!\!\!\perp \Big(S_i(0), S_i(1), Y_i(0, S_i(0)), Y_i(1, S_i(1))\Big) \;\Big|\; X_i.$$

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

Assumption 6.

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

The key assumption we add here, which is not commonly made in the mediation literature, is to rule 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).

The following proposition links the two sets of assumptions.

Proposition 3. *Suppose Assumptions 5-7 hold. Then Assumptions 1 and 2 hold.*

Missing Data. From a missing data perspective, the surrogacy and comparability assumptions we make 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, 2012; Graham, Pinto and Egel 2016).

In our two sample setting, we can think of the complete data as the quintuple $(X_i, S_i, Y_i, W_i, P_i)$. 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 $(X_i, S_i, \mathbf{1}_{P_i=\text{O}} Y_i, \mathbf{1}_{P_i=\text{E}} W_i, P_i)$, where for units with $P_i = \text{O}$ the treatment indicator W_i is missing, and for units with $P_i = \text{E}$ the outcome Y_i is missing. Now consider the following assumption.

Assumption 8. (Missing Data 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 missing at random assumption 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 2 implies surrogacy*

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

and 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 still need the unconfoundedness assumption.

B. Proofs

Proof of Proposition 1:

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

which proves the result. □

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

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

$$\mathbb{E}[Y_i(1)|P_i = \mathrm{E}] = \mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i) \cdot \frac{W_i}{e(X_i)} \middle| P_i = \mathrm{E}\right] \tag{8.1}$$

$$= \mathbb{E}\left[Y_i \cdot \frac{r(S_i, X_i) \cdot t(S_i, X_i) \cdot (1-q)}{e(X_i) \cdot (1 - t(S_i, X_i)) \cdot q} \middle| P_i = \mathrm{O}\right] \tag{8.2}$$

$$= \mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i) \cdot \frac{r(S_i, X_i) \cdot t(S_i, X_i) \cdot (1-q)}{e(X_i) \cdot (1 - t(S_i, X_i)) \cdot q} \middle| P_i = \mathrm{O}\right] \tag{8.3}$$

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 (8.1) uses Assumptions 1–3 and equalities (8.2) and (8.3) only use Assumption 4.

Consider (8.1). By Assumption 1 (unconfoundedness), it follows that

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

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

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

By Assumption 2 (surrogacy), we have

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

By Assumption 3 (comparability), $h_{\mathrm{E}}(s, x) = h_{\mathrm{O}}(s, x)$ so that this is equal to

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

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

Consider (8.2). By the definition of $t(s, x)$, we have

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

where Assumption 4 assures $1 - t(s, x)$ is not zero. This leads to

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

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

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

Using the definition of conditional expectations, we obtain

$$\begin{aligned}
&\mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i) \frac{r(S_i, X_i)}{e(X_i)} \cdot \frac{\mathrm{pr}\left(S_i, X_i \middle| P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i, X_i \middle| P_i = \mathrm{O}\right)} \middle| P_i = \mathrm{O}\right] \\
=& \int h_{\mathrm{O}}(s, x) \frac{r(s, x)}{e(x)} \cdot \frac{\mathrm{pr}\left(S_i = s, X_i = x \middle| P_i = \mathrm{E}\right)}{\mathrm{pr}\left(S_i = s, X_i = x \middle| P_i = \mathrm{O}\right)} \cdot \mathrm{pr}\left(S_i = s, X_i = x \middle| P_i = \mathrm{O}\right) ds dx \\
=& \int h_{\mathrm{O}}(s, x) \frac{r(s, x)}{e(x)} \mathrm{pr}\left(S_i = s, X_i = x \middle| P_i = \mathrm{E}\right) dy ds dx \\
=& \mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i) \frac{r(S_i, X_i)}{e(X_i)} \middle| P_i = \mathrm{E}\right]
\end{aligned}$$

Consider (8.3). By the law of iterated expectations conditional on S_i and X_i, we obtain

$$\mathbb{E}\left[Y_i \cdot \frac{r(S_i, X_i) \cdot t(S_i, X_i) \cdot (1-q)}{e(X_i) \cdot (1 - t(S_i, X_i)) \cdot q} \middle| P_i = \mathrm{O}\right] = \mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i) \cdot \frac{r(S_i, X_i) \cdot t(S_i, X_i) \cdot (1-q)}{e(X_i) \cdot (1 - t(S_i, X_i)) \cdot q} \middle| P_i = \mathrm{O}\right]$$

where Assumption 4 assures $1 - t(s, x)$ is not zero. □

Proof of Theorem 2: Consider part (i). By the law of iterated expectations conditional on S_i and X_i, we have

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

By the proof of (8.2) in Theorem 1 where we don't use surrogacy or comparability, we get

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

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

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

By Assumption 1 (unconfoundedness), we have

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

Undoing the law of iterated expectations give the desired result.

For parts (ii)-(iv), we prove (iv) first. By Assumption 1 (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

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

Thus,we have

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

We add and subtract

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

to get

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

Rearranging the terms, we have

$$\begin{aligned}
&\tau - \mathbb{E}\left[h_{\mathrm{O}}(S_i(1), X_i) - h_{\mathrm{O}}(S_i(0), X_i) \mid P_i = \mathrm{E}\right] && (8.4)\\
=&\mathbb{E}\left[\mathbb{E}\left[\mu_{\mathrm{E}}(S_i, X_i, 1)|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{E}}(S_i, X_i)|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] && (8.5)\\
&- \mathbb{E}\left[\mathbb{E}\left[\mu_{\mathrm{E}}(S_i, X_i, 0)|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] + \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{E}}(S_i, X_i)|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] && (8.6)\\
&+ \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{E}}(S_i, X_i)|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i)|W_i = 1, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] && (8.7)\\
&+ \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{O}}(S_i, X_i)|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] - \mathbb{E}\left[\mathbb{E}\left[h_{\mathrm{E}}(S_i, X_i)|W_i = 0, X_i, P_i = \mathrm{E}\right] \mid P_i = \mathrm{E}\right] && (8.8)
\end{aligned}$$

Next, by the definition of expectations,

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

Use this to write (8.5) as

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

Using the same argument we can write (8.6) as

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

Combining the results for (8.5) and (8.6) leads to

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

Collecting the last two terms, (8.7) and (8.8), we have

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

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

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

Finally for part (ii), under Assumption 2 (surrogacy), but not Assumption 3 (comparability), $\mu_{\mathrm{E}}(S_i,X_i,1)-\mu_{\mathrm{E}}(S_i,X_i,0)=0$ and the result is immediate from (iv). For part (iii), under Assumption 3 (comparability), but not Assumption 2 (surrogacy), $h_{\mathrm{E}}(S_i,X_i)-h_{\mathrm{O}}(S_i,X_i)=0$ and the result is immediate from (iv). □

Proof of Theorem 3: The first representation of the efficiency bound without surrogacy is derived in Robins and Rotnitzky (1995); Robins, Rotnitzky and Zhao (1995); Hahn (1998). For the second case we focus on the setting where the propensity score is constant, and the surrogate is discrete with support $s_1,\ldots,s_M$. The latter is not restrictive, and the former can be relaxed at the expense of additional algebra.

The efficient estimator is $\hat{\tau}=\overline{Y}_1-\overline{Y}_0$ where $\overline{Y}_1$ and $\overline{Y}_0$ are the average values for the surrogate outcome in treated and control samples respectively. We can write this as

$$
\hat{\tau}=\sum_{m=1}^{M}\hat{\pi}_{s|1}\cdot\hat{\mu}_{\mathrm{E}}(s_m,1)-\sum_{m=1}^{M}\hat{\pi}_{s|0}\cdot\hat{\mu}_{\mathrm{E}}(s_m,0).
$$

Here $\hat{\mu}_{\mathrm{E}}(s,w)$ is the average outcome for units with $S_i=s$ and $W_i=w$, and $\hat{\pi}_{\mathrm{E}}(s|w)=P(S_i=s|W_i=w)$. Let $\hat{\pi}_{\mathrm{E}}(s)$ be the fraction of units with $S_i=s$. Let $\pi_{\mathrm{E}}(s|w)$ and $\pi_{\mathrm{E}}(s)$ be the corresponding population probabilities, so that $\pi_{\mathrm{E}}(s|1)=\pi_{\mathrm{E}}(s)\cdot r(s)/p$.

We can write the difference between $\hat{\tau}$ and $\tau = \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|1) \cdot \mu_{\mathrm{E}}(s_m, 1) - \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|0) \cdot \mu(s_m, 0)$ as

$$\hat{\tau} - \tau = \sum_{m=1}^{M} \hat{\pi}_{\mathrm{E}}(s_m|1) \cdot (\hat{\mu}_{\mathrm{E}}(s_m, 1) - \mu_{\mathrm{E}}(s_m, 1)) - \sum_{m=1}^{M} \hat{\pi}_{\mathrm{E}}(s_m|0) \cdot (\hat{\mu}_{\mathrm{E}}(s_m, 0) - \mu_{\mathrm{E}}(s_m, 0))$$

$$+ \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s_m|1) - \pi_{\mathrm{E}}(s_m|1)) \cdot \mu_{\mathrm{E}}(s_m, 1) - \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s_m|0) - \pi_{\mathrm{E}}(s_m|0)) \cdot \mu_{\mathrm{E}}(s_m, 0).$$

Up to the relevant order of approximation this is equal to

$$\hat{\tau} - \tau \approx \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|1) \cdot (\hat{\mu}_{\mathrm{E}}(s_m, 1) - \mu_{\mathrm{E}}(s_m, 1)) - \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|0) \cdot (\hat{\mu}_{\mathrm{E}}(s_m, 0) - \mu_{\mathrm{E}}(s_m, 0))$$

$$+ \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s|1) - \pi_{\mathrm{E}}(s|1)) \cdot \mu_{\mathrm{E}}(s_m, 1) - \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s|1) - \pi_{\mathrm{E}}(s|0)) \cdot \mu_{\mathrm{E}}(s_m, 0).$$

If N is the overall sample size, the variance of $\hat{\mu}_{\mathrm{E}}(s, 1)$ is $\sigma^2(s)/(N \cdot \pi_{\mathrm{E}}(s|1) \cdot p)$. The variance of $\hat{\pi}_{\mathrm{E}}(s|1)$ is $\pi_{\mathrm{E}}(s|1) \cdot (1 - \pi_{\mathrm{E}}(s|1))/(N \cdot p)$. Then

$$\mathbb{V}_{\mathrm{E}}(\hat{\tau} - \tau) \approx \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|1)^2 \cdot \frac{\sigma^2(s_m)}{N \cdot \pi_{\mathrm{E}}(s_m|1) \cdot p} - \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m|0)^2 \cdot \frac{\sigma^2(s_m)}{N \cdot \pi_{\mathrm{E}}(s_m|0) \cdot (1-p)}$$

$$+ \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m|1) \cdot (1 - \pi_{\mathrm{E}}(s_m|1))}{N \cdot p} \cdot (\mu_{\mathrm{E}}(s_m, 1) - \mu_1)^2 + \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m|0) \cdot (1 - \pi_{\mathrm{E}}(s_m|0))}{N \cdot (1-p)} \cdot (\mu_{\mathrm{E}}(s_m, 0) - \mu_0)^2$$

$$\approx \sum_{m=1}^{M} r(s_m) \cdot \pi_{\mathrm{E}}(s_m) \cdot \frac{\sigma^2(s_m)}{N \cdot p^2} - \sum_{m=1}^{M} r(s_m) \cdot \pi_{\mathrm{E}}(s_m) \frac{\sigma^2(s_m)}{N \cdot (1-p)^2}$$

$$+ \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m|1)}{N \cdot p} \cdot (\mu_{\mathrm{E}}(s_m, 1) - \mu_1)^2 + \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m|0)}{N \cdot (1-p)} \cdot (\mu_{\mathrm{E}}(s_m, 0) - \mu_0)^2$$

$$\sum_{m=1}^{M} r(s_m) \cdot \pi_{\mathrm{E}}(s_m) \cdot \frac{\sigma^2(s_m)}{N \cdot p^2} - \sum_{m=1}^{M} r(s_m) \cdot \pi_{\mathrm{E}}(s_m) \frac{\sigma^2(s_m)}{N \cdot (1-p)^2}$$

$$+ \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m) \cdot r(s_m)}{N \cdot p^2} \cdot (\mu_{\mathrm{E}}(s_m, 1) - \mu_1)^2 + \sum_{m=1}^{M} \frac{\pi_{\mathrm{E}}(s_m) \cdot r(s_m)}{N \cdot (1-p)^2} \cdot (\mu_{\mathrm{E}}(s_m, 0) - \mu_0)^2$$

$$= \frac{1}{N} \cdot \mathbb{E}_{\mathrm{E}}\Bigg[\sigma^2(S_i) \cdot \left(\frac{r(S_i)}{p^2} + \frac{1 - r(S_i)}{(1-p)^2}\right)$$

$$+ \frac{r(S_i)}{p^2} \cdot (\mu(S_i) - \mu_1)^2 + \frac{1 - r(S_i)}{(1-p)^2} \cdot (\mu(S_i) - \mu_0)^2\Bigg].$$

Now consider the case with surrogacy. The estimator now is

$$\hat{\tau} - \tau = \sum_{m=1}^{M} \hat{\pi}_{\mathrm{E}}(s_m|1) \cdot (\hat{h}_{\mathrm{E}}(s_m) - \mu_{\mathrm{E}}(s_m, 1)) - \sum_{m=1}^{M} \hat{\pi}_{\mathrm{E}}(s|0) \cdot (\hat{h}_{\mathrm{E}}(s_m) - \mu_{\mathrm{E}}(s_m, 0))$$
$$+ \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s|1) - \pi_{\mathrm{E}}(s|1)) \cdot \mu_{\mathrm{E}}(s_m, 1)) - \sum_{m=1}^{M} (\hat{\pi}_{\mathrm{E}}(s|0) - \pi_{\mathrm{E}}(s|0) \cdot \mu_{\mathrm{E}}(s_m, 0)),$$

where $\hat{h}_{\mathrm{E}}(s)$ is the average outcome for all units with $S_i = s$, no longer separately by treatment status. Approximately, the estimator is

$$\hat{\tau} - \tau = \sum_{m=1}^{M} \pi_{s|1} \cdot (\hat{h}_{\mathrm{E}}(s_m) - \mu(s_m, 1)) - \sum_{m=1}^{M} \pi_{s|0} \cdot (\hat{h}_{\mathrm{E}}(s_m) - \mu(s_m, 0))$$
$$+ \sum_{m=1}^{M} (\hat{\pi}_{s|1} - \pi_{s|1}) \cdot \mu(s_m, 1)) - \sum_{m=1}^{M} (\hat{\pi}_{s|0} - \pi_{s|0}) \cdot \mu(s_m, 0)).$$

The variance for the last two terms does not change, but the variance for the first two terms is different, and there is also a covariance term. The total variance of the first term is

$$\sum_{m=1}^{M} (\pi_{\mathrm{E}}(s_m|1) - \pi_{\mathrm{E}}(s_m|0))^2 \cdot \mathbb{V}(\hat{h}_{\mathrm{E}}(s_m))$$
$$= \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m)^2 \left(\frac{r(s_m)}{p} - \frac{1 - r(s_m)}{1 - p} \right)^2 \cdot \frac{\sigma^2(s_m)}{N \cdot \pi_{\mathrm{E}}(s_m)}$$
$$= \frac{1}{N} \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m) \left(\frac{r(s_m) - p}{p \cdot (1-p)} \right)^2 \cdot \sigma^2(s_m)$$
$$= \frac{1}{N} \sum_{m=1}^{M} \pi_{\mathrm{E}}(s_m) \left(\frac{r(s_m)}{p} + \frac{1 - r(s_m)}{1-p} - \frac{r(s_m) \cdot (1 - r(s_m))}{p^2 \cdot (1-p)^2} \right) \cdot \sigma^2(s_m)$$
$$= \frac{1}{N} \cdot \mathbb{E}_{\mathrm{E}} \left[\sigma^2(S_i) \cdot \left(\frac{r(S_i)}{p^2} + \frac{1 - r(S_i)}{(1-p)^2} - \frac{r(S_i) \cdot (1 - r(S_i))}{p^2 \cdot (1-p)^2} \right) \right.$$

Combining this with the last term leads to

$$\mathbb{V}_{\mathrm{E}}(\hat{\tau}) \approx \frac{1}{N} \cdot \mathbb{E}_{\mathrm{E}} \left[\sigma^2(S_i) \cdot \left(\frac{r(S_i)}{p^2} + \frac{1 - r(S_i)}{(1-p)^2} - \frac{r(S_i) \cdot (1 - r(S_i))}{p^2 \cdot (1-p)^2} \right) \right.$$
$$\left. + \frac{r(S_i)}{p^2} \cdot (h_{\mathrm{E}}(S_i) - \mu_1)^2 + \frac{1 - r(S_i)}{(1-p)^2} \cdot (h_{\mathrm{E}}(S_i) - \mu_0)^2 \right].$$

□

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

$$W_i \perp\!\!\!\perp \Big(S_i(0), S_i(1), Y_i(S_i(0)), Y_i(1, S_i(1)) \Big) \;\Big|\; X_i \tag{8.9}$$

$$S_i \perp\!\!\!\perp \Big(Y_i(W_i, s)_{s\in\mathbb{S}}\Big) \;\Big|\; X_i, W_i \tag{8.10}$$

and

$$Y_i(w, s) = Y_i(w', s) \qquad \forall\ i, w, w' \in \mathbb{W}, s \in \mathbb{S}, \tag{8.11}$$

imply

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1), S_i(0), S_i(1)\Big) \;\Big|\; X_i, \tag{8.12}$$

$$W_i \perp\!\!\!\perp Y_i \;\Big|\; S_i, X_i. \tag{8.13}$$

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 (8.12) follows directly from (8.9) because $Y_i(w) = Y_i(w, S_i(w))$.

Condition (8.11) implies that we can write $Y_i(s)$ without ambiguity, and by (8.9), we have

$$W_i \perp\!\!\!\perp Y_i(s) \;\Big|\; X_i.$$

By (8.10) 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 estimate from the data the distributions

$$f_{Y|S,X,P}(y|s,x,\mathrm{O}), \qquad f_{W|S,X,P}(w|s,x,\mathrm{E}), \qquad \text{and} \quad f_{P,S,X}(p,s,x),$$

but no other distributions. That implies that the joint distribution of (Y,W,P,S,X) implied by

$$f_{Y|W,S,X,P}(y|w,s,x,p) = f_{Y|S,X,P}(y|s,x,\mathrm{O}),$$

and

$$f_{W|S,X,P}(w|s,x,\mathrm{O}) = f_{W|S,X,P}(w|s,x,\mathrm{E}),$$

for all (y,w,p,s,x) is consistent with the data, and it also satisfies Assumption 8. □

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.

Abadie, Alberto, and Matias D Cattaneo. 2018. "Econometric methods for program evaluation." *Annual Review of Economics*, 10: 465–503.

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.

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters*, 178: 58–62.

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: 444–472.

Athey, Susan, and Scott Stern. 1998. "An empirical framework for testing theories about complimentarity in organizational design." National Bureau of Economic Research.

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.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 28(2): 29–50.

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.

Breiman, Leo. 2001. "Random forests." *Machine Learning*, 45(1): 5–32.

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, 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.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz. 2016. "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment." *American Economic Review*, 106(4): 855–902.

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 pointsï¿œapplication to comparison of different breast screening frequencies." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(1): 49–60.

Ding, Peng, and Jiannan Lu. 2017. "Principal stratification analysis using principal scores." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 757–777.

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.

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

Hansen, Ben B. 2008. "The prognostic analogue of the propensity score." *Biometrika*, 95(2): 481–488.

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.

Hendren, Nathaniel, and Benjamin D Sprung-Keyser. 2019. "A Unified Welfare Analysis of Government Policies." National Bureau of Economic Research.

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 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.

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

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 Jeffrey M Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47(1): 5–86.

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.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions*." *The Quarterly Journal of Economics*, 114(2): 497–532.

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.

Mealli, Fabrizia, and Alessandra Mattei. 2012. "A refreshing account of principal stratification." *The international journal of biostatistics*, 8(1).

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.

Oster, Emily. 2019. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

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

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

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. "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. 2006. *Matched sampling for causal effects.* Cambridge University Press.

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

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

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

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

Van der Laan, Mark J, and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.

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

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

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.
document

TABLE 1
Setup and Notation

(✓ 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_{E}	E	✓	✗	✓	✓
$N_{\mathrm{E}}+1$ to $N_{\mathrm{E}}+N_{\mathrm{O}}$	O	✗	✓	✓	✓

Notes: This table summarizes the setup of the dataset for the problem we analyze. The experimental sample has N_{E} observations. It includes information on the treatment indicator, the intermediate outcomes (surrogates), and pretreatment variables; however, critically, it does not include data on the long-term outcome, reflecting the fact that the long-term outcome is observed with a long delay after the experiment. The observational sample has N_{O} observations. It includes information on the primary outcome, the surrogates, and the pretreatment variables; however, it may not include data on the treatment variable. The question we analyze is how one can identify the causal effect of W on Y in this setting where W and Y are not observed together in the experimental sample.

APPENDIX TABLE 1

Regression Coefficients Underlying Surrogate Indices for Mean Employment and Earnings Over Nine Years

	Employment		Earnings	
Quarter	*Six-Quarter Surrogate Index (1)*	*Twelve-Quarter Surrogate Index (2)*	*Six-Quarter Surrogate Index (3)*	*Twelve-Quarter Surrogate Index (4)*
1	0.062	0.043	0.078	-0.007
	(0.010)	(0.008)	(0.021)	(0.015)
2	0.036	0.033	-0.040	-0.015
	(0.011)	(0.008)	(0.021)	(0.015)
3	0.072	0.042	0.093	0.055
	(0.011)	(0.009)	(0.021)	(0.015)
4	0.054	0.030	0.057	0.012
	(0.012)	(0.009)	(0.021)	(0.015)
5	0.113	0.052	0.080	0.046
	(0.012)	(0.009)	(0.021)	(0.015)
6	0.217	0.045	0.390	0.045
	(0.011)	(0.010)	(0.016)	(0.015)
7		0.042		0.062
		(0.010)		(0.016)
8		0.037		0.009
		(0.010)		(0.016)
9		0.073		0.079
		(0.010)		(0.016)
10		0.039		0.041
		(0.011)		(0.016)
11		0.095		0.115
		(0.011)		(0.017)
12		0.204		0.315
		(0.009)		(0.013)
Constant	0.131	0.072	453.143	250.705
	(0.005)	(0.004)	(21.825)	(15.980)
Estimated Treatment Effect	0.067	0.074	249.306	277.159
	(0.007)	(0.009)	(36.340)	(46.706)

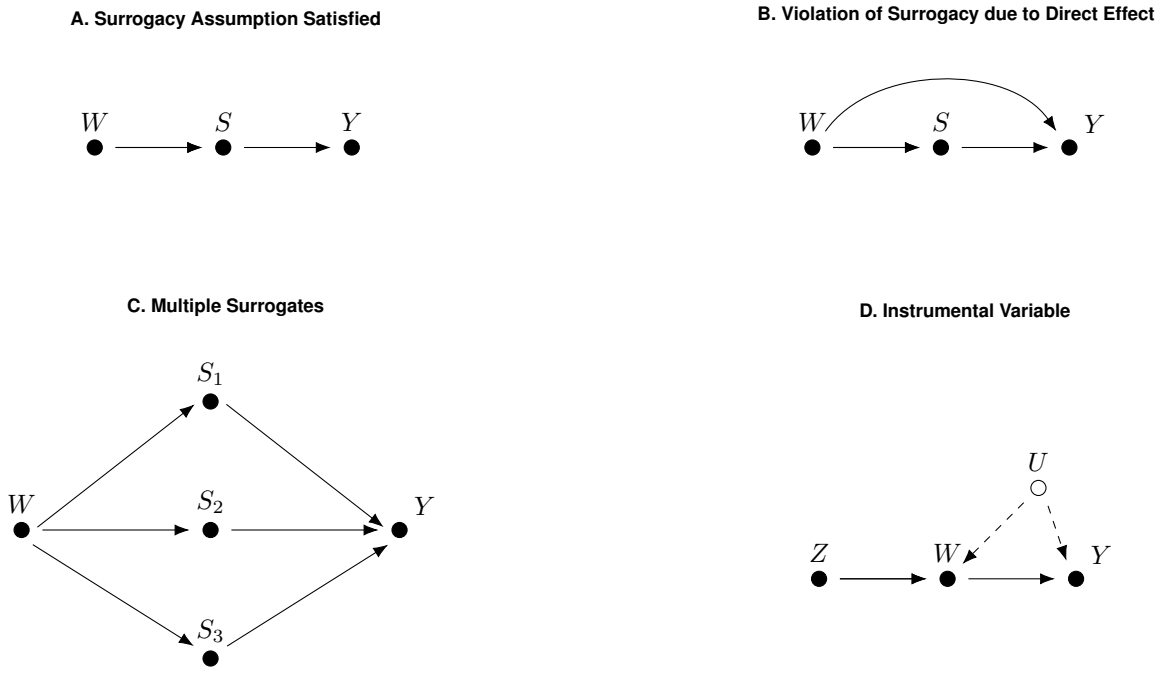
Notes: This table shows the regression coefficients on employment (Columns 1 and 2) and earnings (Columns 3 and 4) used to construct surrogate indices for mean employment and earnings over nine years. These are simply regression coefficient of the nine-year mean outcome on employment rates or earnings over the first six quarters (Columns 1 and 3) or twelve quarters (Columns 2 and 4). Standard errors for the coefficients are shown in parentheses. The estimated treatment effects listed in the final row show the effect on mean employment or earnings over 36 quarters estimated using the surrogate index.

APPENDIX TABLE 2
Estimates of Treatment Effects on Employment and Earnings Over Nine Years, Varying Quarters of Data Used to Construct Surrogate Index

Quarters Used	*Employment* (1)	*Earnings* (2)
1	0.012	53.589
	(0.003)	(16.705)
2	0.032	125.402
	(0.004)	(20.981)
3	0.046	166.467
	(0.005)	(26.858)
4	0.053	193.449
	(0.006)	(29.976)
5	0.063	221.958
	(0.006)	(33.018)
6	0.067	249.306
	(0.007)	(36.340)
7	0.067	244.448
	(0.007)	(38.754)
8	0.071	260.034
	(0.008)	(40.296)
9	0.075	287.194
	(0.008)	(42.131)
10	0.075	284.252
	(0.008)	(43.490)
11	0.075	283.595
	(0.008)	(45.171)
12	0.074	277.159
	(0.009)	(46.706)
13	0.072	273.788
	(0.009)	(47.550)
14	0.074	288.479
	(0.009)	(48.458)
15	0.074	284.172
	(0.009)	(49.152)
16	0.071	283.263
	(0.009)	(49.719)
17	0.073	294.511
	(0.010)	(50.609)
18	0.073	290.171
	(0.010)	(51.389)
19	0.072	293.753
	(0.010)	(52.130)
20	0.074	290.306
	(0.010)	(52.615)
21	0.073	285.737
	(0.010)	(52.995)
22	0.072	286.378
	(0.010)	(53.465)
23	0.072	286.091
	(0.010)	(53.823)
24	0.071	269.948
	(0.010)	(54.297)
25	0.071	274.255
	(0.010)	(54.617)
26	0.068	263.414
	(0.010)	(55.126)
27	0.067	256.391
	(0.010)	(55.361)
28	0.065	254.838
	(0.010)	(55.535)
29	0.066	261.461
	(0.011)	(55.707)
30	0.066	260.665
	(0.011)	(55.840)
31	0.066	255.566
	(0.011)	(55.963)
32	0.066	256.100
	(0.011)	(56.034)
33	0.065	252.594
	(0.011)	(56.125)
34	0.064	250.590
	(0.011)	(56.164)
35	0.064	249.931
	(0.011)	(56.182)
36	0.064	249.054
	(0.011)	(56.210)

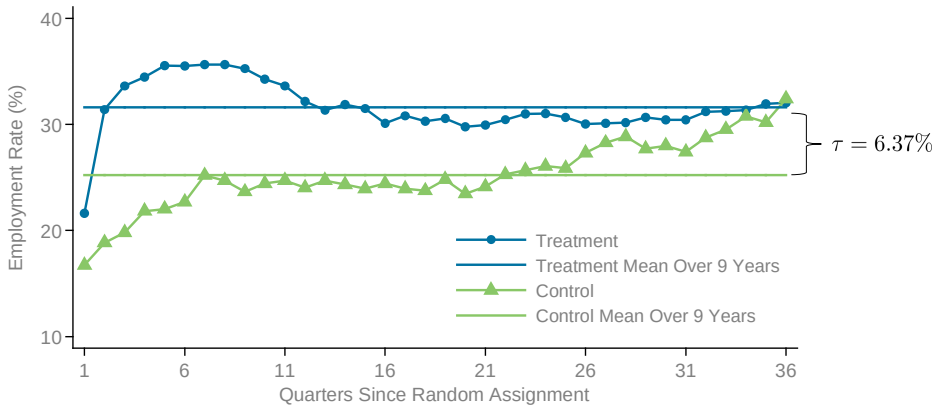
Notes: Column 1 shows estimated treatment effects on a surrogate index for mean employment rates over 9 years that is constructed using data on employment rates up to quarter x, varying x from 1 to 36. The surrogate index is constructed as the predicted value from an individual-level OLS regression of mean employment rates over 36 quarters on employment indicators from the first quarter after the intervention to quarter x (using data from the treatment group only to fit the model). The point estimates that are listed exactly match those plotted in Figure 3a; standard errors for these estimates are given in parentheses. Column 2 replicates Column 1 using mean quarterly earnings rather than employment as the outcome variable, and using mean quarterly earnings rather than employment to construct surrogate indices, as in Appendix Figure 2a.

FIGURE 1
Surrogacy Assumptions and Violations



Notes: This figure shows a set of directed acyclical graphs (DAGs) that illustrate the assumptions underlying the use of an intermediate outcome S as a surrogate to estimate the causal effect of a treatment W on a long-term outcome Y. Panel A shows a case where the surrogacy assumption is satisfied with a single intermediate outcome S which lies on the causal chain from W to Y. Panel B shows a case where this surrogacy condition is violated, due to a direct effect of the treatment on the outcome that doesn't pass through the surrogate. Panel C shows how this problem can be addressed by using multiple intermediate outcomes to capture other pathways that link the treatment and long-term outcome, which (in combination with unconfoundedness and comparability assumptions) is the logic underlying our surrogate index approach. Finally, in Panel D, we present a DAG representation of standard instrumental variables estimators, where there is an unobserved confound between W and Y that is addressed by introducing an instrument Z that affects W but does not affect Y directly (the exclusion restriction). The instrumental variables case differs from the case on which we center our attention here because we focus on a setting in which W and Y are not observed in the same sample, as shown in Table 1.

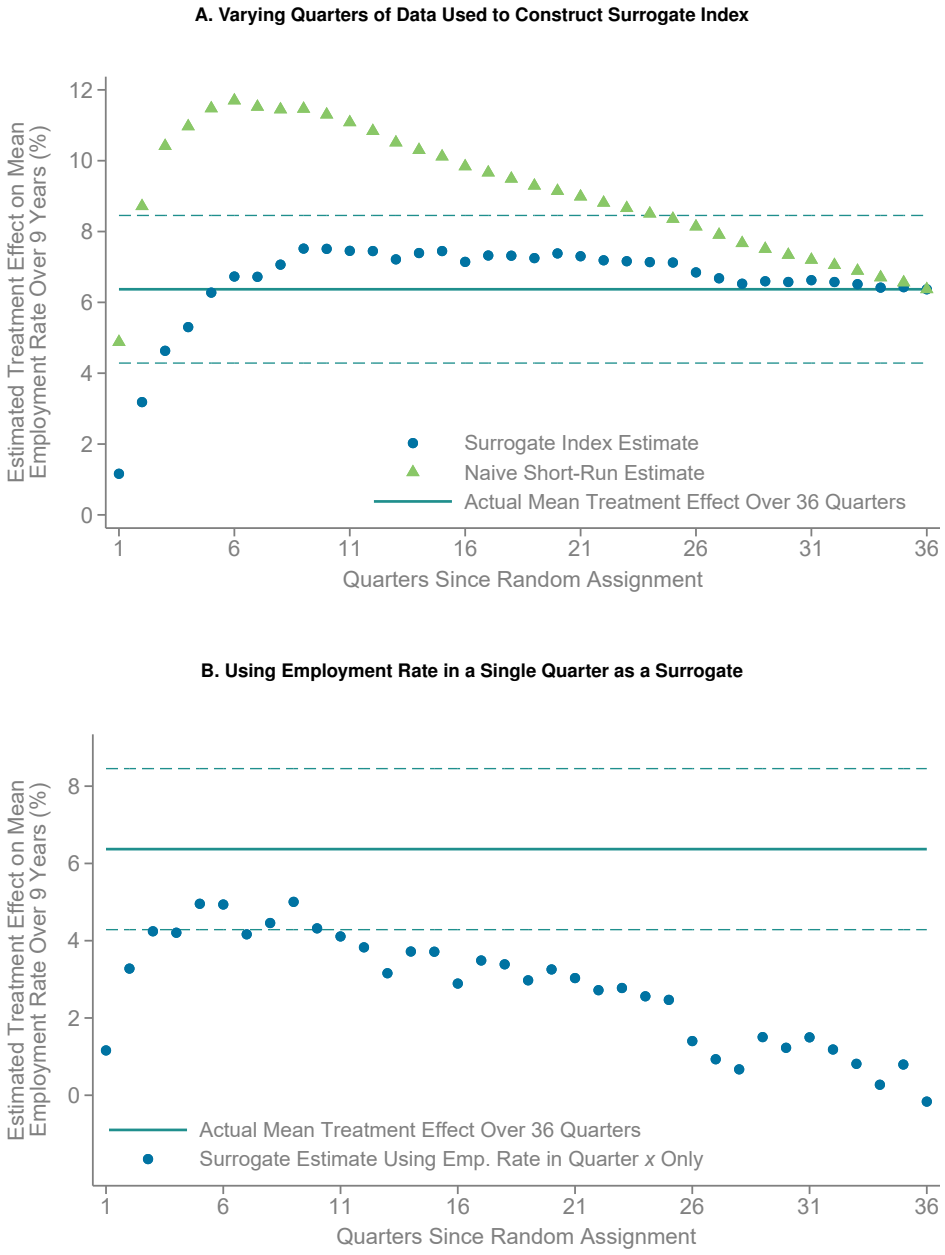
FIGURE 2
Employment Rates in Riverside GAIN Treatment vs. Control Group, by Quarter



Notes: This figure shows the employment rates for the treatment group (circles) and control group (triangles) in the California GAIN training program in Riverside County, CA. Each point represents the employment rate among the group at a given number of quarters since random assignment. The horizontal lines show the mean employment rate over the nine years (36 quarters) after random assignment in the two groups; the difference between these two groups is the long-term treatment effect τ of interest. The data underlying this figure, which are based on Unemployment Insurance records, were obtained from Hotz, Imbens and Klerman (2006).

FIGURE 3

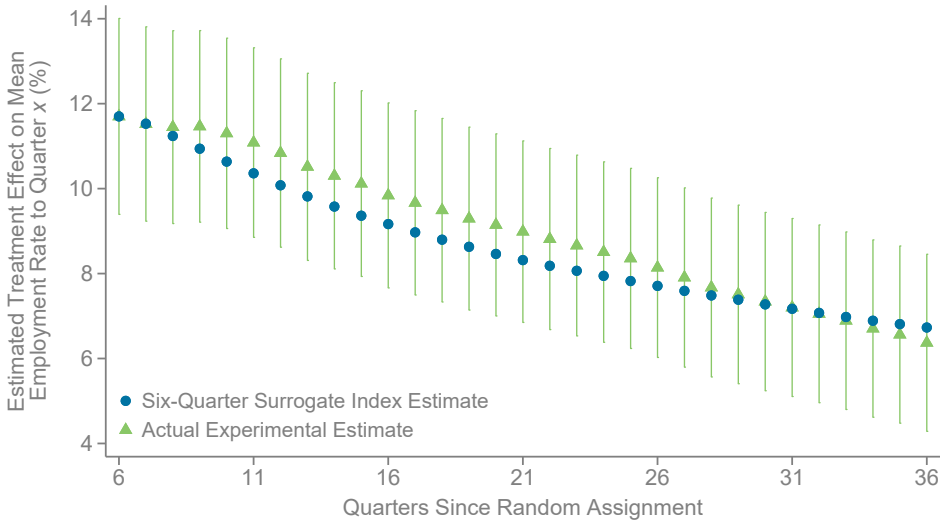
Estimates of Treatment Effect on Mean Employment Rates Over Nine Years



Notes: These figures show how surrogate indices based on varying numbers of quarters of employment data perform in matching the actual treatment effect of the GAIN program on mean employment rates over nine years in Riverside. In both panels, the solid line shows the mean treatment effect of $\tau = 6.37\%$ on employment rates over nine years (36 quarters), estimated as the difference in means between the treatment and control groups as shown in Figure 2. The dashed lines show a 95% confidence interval for this estimate. In Panel A, the circles show the estimated treatment effect on a surrogate index for mean employment rates over 9 years that is constructed using data on employment rates up to quarter x, varying x from 1 to 36. The surrogate index is constructed as the predicted value from an individual-level OLS regression of mean employment rates over 36 quarters on employment indicators from the first quarter after the intervention to quarter x (using data from the treatment group only to fit the model). The triangles in Panel A show a "naive" estimate of the mean treatment effect on observed employment rates up to quarter x. In Panel B, the circles show the treatment effect estimate obtained if one uses *only* the employment indicator in quarter x as the surrogate instead of constructing an index based on employment rates from quarter 1 to quarter x. This series is constructed in the same way as the series in circles in Panel A, except that we use only the employment indicator in quarter x as a predictor when constructing the surrogate index.

FIGURE 4

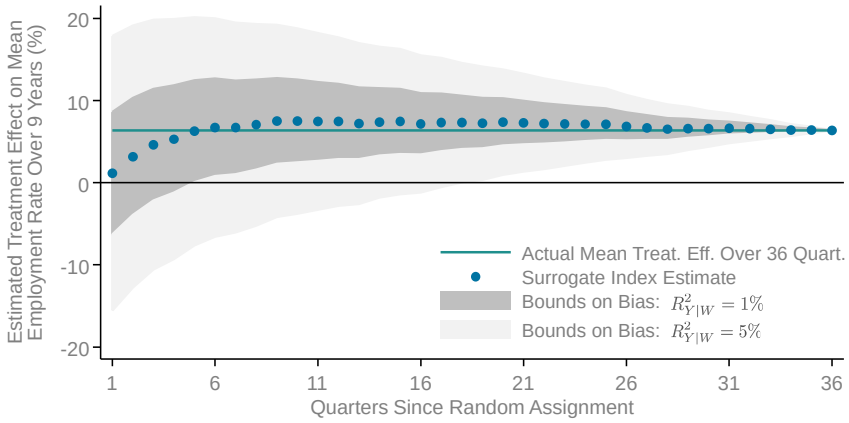
Validation of Six-Quarter Surrogate Index: Estimates of Treatment Effects on Mean Employment Rates, Varying Outcome Horizon



Notes: This figure shows how a surrogate index based on the first six quarters of employment data can be validated out of sample by comparing estimates based on the surrogate index to actual observed treatment effects on employment outcomes over time in Riverside. The triangles show point estimates of the treatment effect on cumulative mean employment rates from quarter 1 to quarter x after random assignment. The solid vertical lines show 95% confidence intervals for these estimates. The circles show estimates of treatment effects based on a surrogate index for mean employment rates up to quarter x constructed using the first six quarters of employment data. The surrogate index is constructed as the predicted value from an individual-level OLS regression of mean employment rates up to quarter x on employment indicators in the first six quarters after random assignment (using data from the treatment group only to fit the model).

FIGURE 5

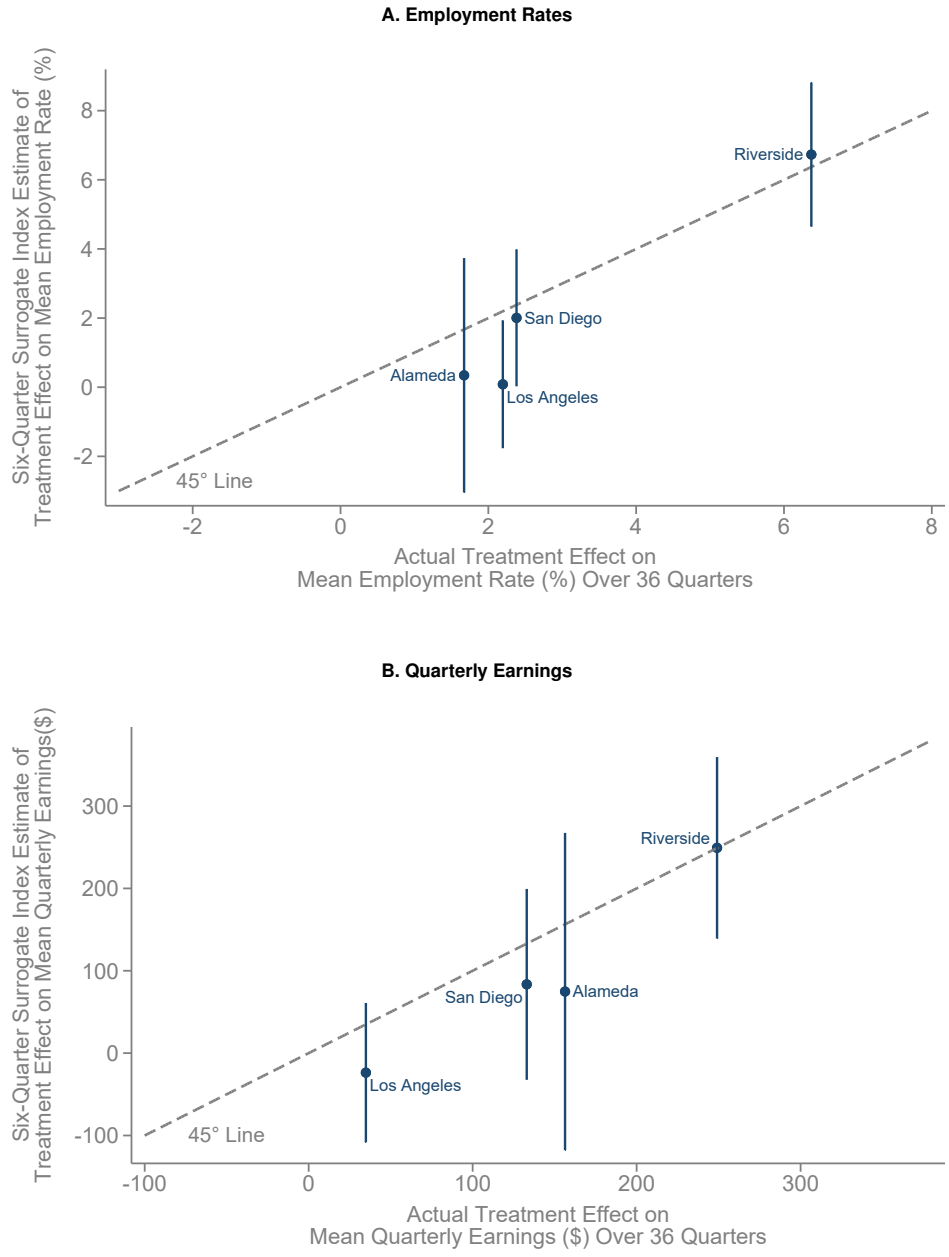
Bounds on Mean Treatment Effect on Employment Rates Over Nine Years, Varying Number of Quarters Used to Construct Surrogate Index



Notes: This figure plots bounds on the degree of bias from violations of the surrogacy assumption when using a surrogate index to estimate the treatment effect on long-term employment rates in Riverside. The dark grey shaded regions shows bounds on the treatment effect on mean employment rates over nine years based on surrogate index estimates constructed using employment data up to quarter x, varying x from 1 to 36, assuming that the treatment explains at most $R^2_{Y|W} = 1\%$ of the variation in mean employment rates. These bounds are calculated using the formula in equation (4.3). The light grey shaded region replicates the dark grey shaded region assuming $R^2_{Y|W} = 5\%$. The horizontal line shows the actual mean treatment effect estimate of $\tau = 6.37\%$ on employment rates over nine years (36 quarters). The circles show the point estimate of the treatment effect using a surrogate index for mean employment rates over 9 years that is constructed using data on employment rates up to quarter x, varying x from 1 to 36. The series in circles replicates the series in circles plotted in Figure 2a; see notes to that figure for further details.

FIGURE 6

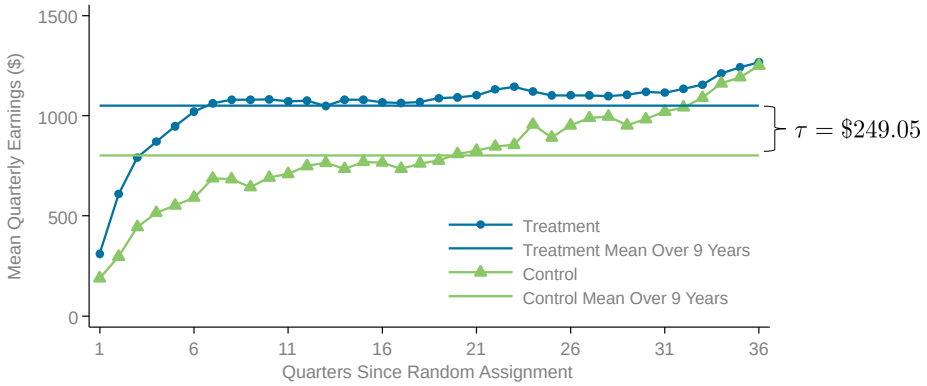
Surrogate Index Estimates vs. Actual Experimental Estimates, by Site



Notes: This figure evaluates the performance of the surrogate index estimated using data from Riverside in predicting heterogeneity in treatment effects across sites. Panel A plots treatment effect estimates based on the six-quarter surrogate index against treatment effect estimates based on actual mean employment rates over nine years. We construct the surrogate index for employment rates by first running an individual-level OLS regression of mean employment rates over 36 quarters on employment indicators from the first to sixth quarter after the intervention, using data from the treatment group in Riverside only. We then use these regression coefficients to predict employment rates for individuals in all sites. Finally, we estimate treatment effects as the mean difference beteween the predicted employment rates for the treatment and control groups in each site. The actual experimental estimates are simply the observed differences in mean employment rates over nine years between the treatment and control groups in each site. The vertical dashed lines show 95% confidence intervals for the actual experimental estimates. The earnings estimates in Panel B are constructed analogously, using quarterly earnings rather than employment as the outcome variable, and using quarterly earnings rather than employment indicators to construct surrogate indices.

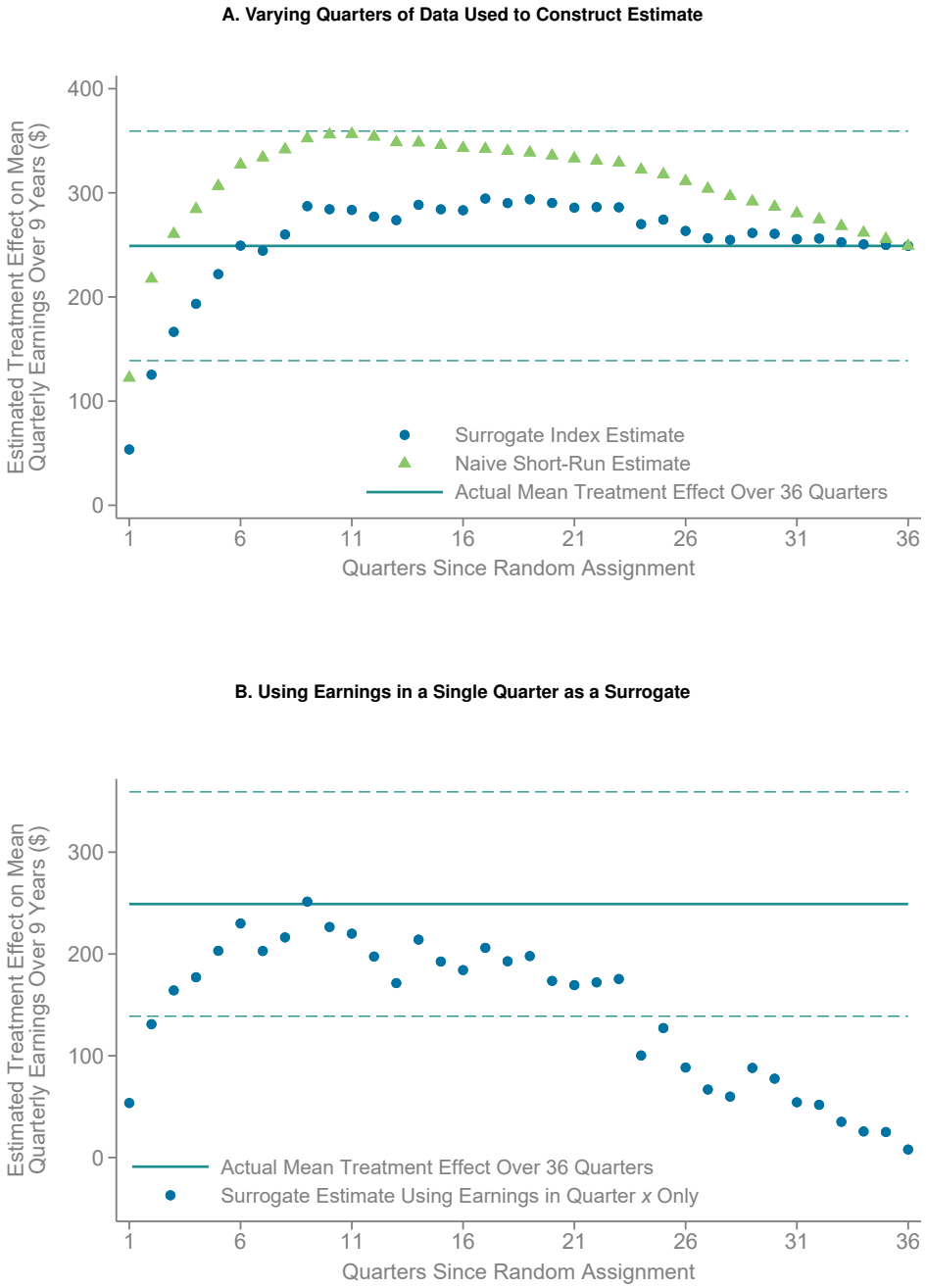
APPENDIX FIGURE 1

Earnings in GAIN Treatment vs. Control Group, by Quarter



Notes: This figure replicates Figure 2 using quarterly earnings rather than employment as the outcome variable. See notes to Figure 2 for details.

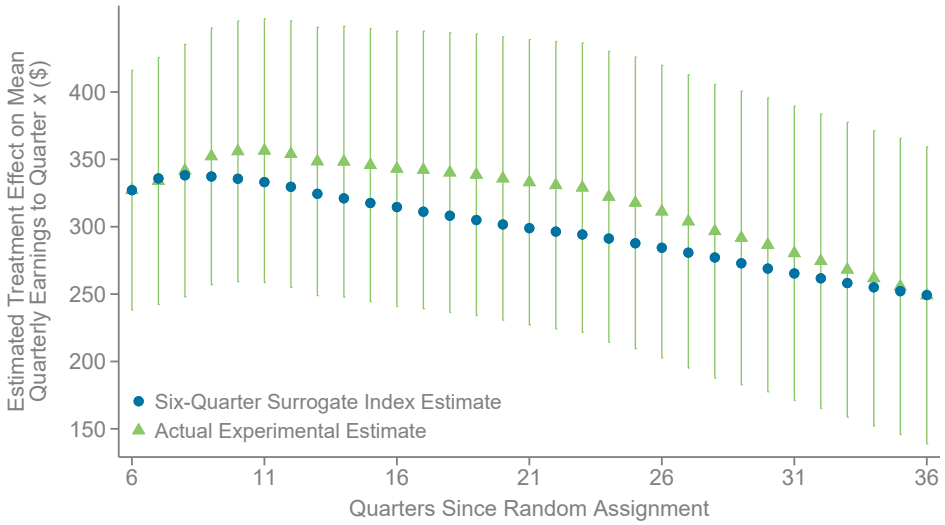
APPENDIX FIGURE 2
Estimates of Treatment Effect on Mean Quarterly Earnings Over Nine Years



Notes: This figure replicates Figure 3 using quarterly earnings rather than employment as the outcome variable, and using quarterly earnings rather than employment indicators to construct surrogate indices. See notes to Figure 3 for details.

APPENDIX FIGURE 3

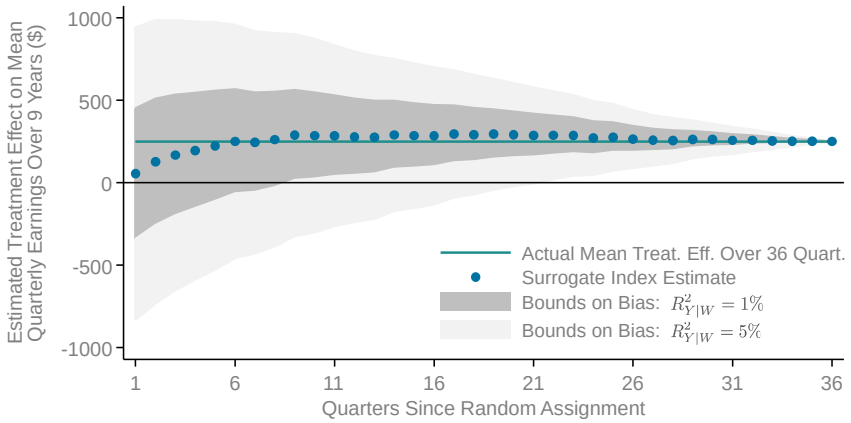
Validation of Six-Quarter Surrogate Index: Estimates of Treatment Effects on Mean Quarterly Earnings, Varying Outcome Horizon



Notes: This figure replicates Figure 4 using quarterly earnings rather than employment as the outcome variable, and using quarterly earnings rather than employment indicators to construct surrogate indices. See notes to Figure 4 for details.

APPENDIX FIGURE 4

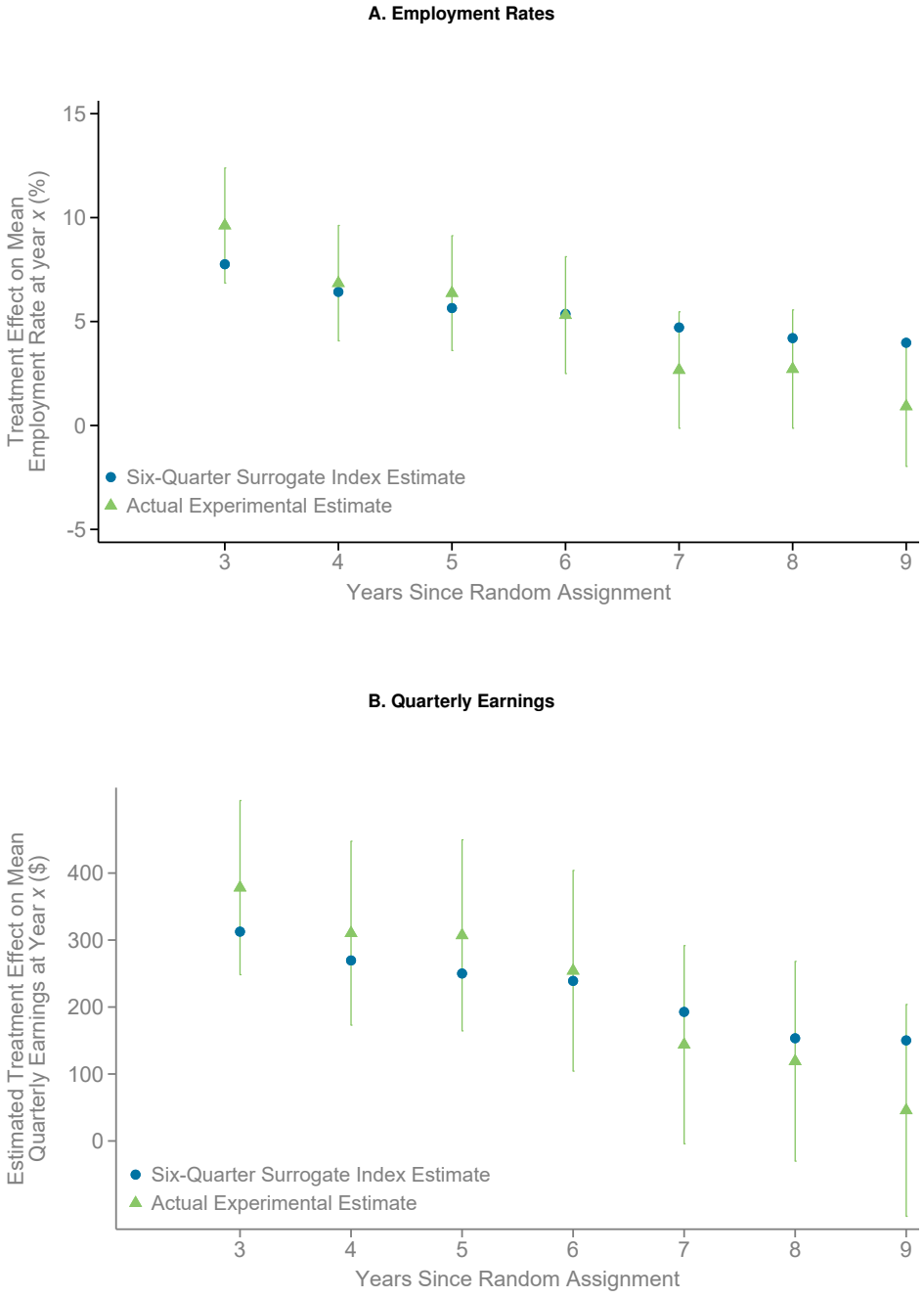
Bounds on Mean Treatment Effect on Earnings Over Nine Years, Varying Number of Quarters Used to Construct Surrogate Index



Notes: This figure replicates Figure 5 using quarterly earnings rather than employment as the outcome variable, and using quarterly earnings rather than employment indicators to construct surrogate indices. See notes to Figure 5 for details.

APPENDIX FIGURE 5

Treatment Effects on Employment and Earnings in Each Year, Varying Outcome Horizon



Notes: This figure replicates Figure 4 and Appendix Figure 3 using yearly means instead of cumulative means of employment rates (Panel A) and quarterly earnings (Panel B). In Panel A, the triangles show the actual experimental estimate (based on the mean difference between the treatment and control group) in each year x after random assignment, varying x from 3 (quarters 8-11) to 9 (quarters 33-36). The vertical lines show 95% confidence intervals for these estimates. The circles show estimates of treatment effects based on a surrogate index for mean employment rates in year x constructed using six quarters of employment data. The surrogate index is constructed as the predicted value from an individual-level OLS regression of mean employment rates in year x on employment indicators in the first six quarters after random assignment (using data from the treatment group only to fit the model). Panel B replicates Panel A using quarterly earnings instead of employment rates as the outcome and for the construction of the surrogate index.