Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index *

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

Current version June 2016

Abstract

Estimating the long-term effects of treatments is of interest in many fields. A common challenge in estimating such treatment effects is that long-term outcomes are unobserved in the time frame needed to make policy decisions. One approach to overcome this missing data problem is to analyze treatments effects on an intermediate outcome, often called a statistical surrogate, if it satisfies the condition that treatment and outcome are independent conditional on the statistical surrogate. The validity of the surrogacy condition is often controversial. Here we exploit that fact that in modern datasets, researchers often observe a large number, possibly hundreds or thousands, of intermediate outcomes, thought to lie on or close to the causal chain between the treatment and the long-term outcome of interest. Even if none of the individual proxies satisfies the statistical surrogacy criterion by itself, using multiple proxies can be useful in causal inference. We focus primarily on a setting with two samples, an experimental sample containing data about the treatment indicator and the surrogates and an observational sample containing information about the surrogates and the primary outcome. We state assumptions under which the average treatment effect be identified and estimated with a high-dimensional vector of proxies that collectively satisfy the surrogacy assumption, and derive the bias from violations of the surrogacy assumption, and show that even if the primary outcome is also observed in the experimental sample, there is still information to be gained from using surrogates.

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

*We are grateful for discussions with Scott Stern, Liang Xu, James Dailey, Dylan Small, and for comments from seminar participants at the University of Pennsylvania, Berkeley, Stanford, and Carnegie Mellon and for financial support from the NSF through grant DMS-1502437.

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

[‡]Department of Economics, Stanford University, and NBER, chetty@stanford.edu.

[§]Graduate School of Business, Stanford University, and NBER, imbens@stanford.edu.

[¶]NSF Mathematical Science Postdoctoral Research Fellow and SIEPR, Stanford University, hskang@stanford.edu.

1 Introduction

Estimating the long-term effects of treatments is of interest in many fields, ranging from medicine (e.g., the effects of drugs on mortality rates) to economics (e.g., the effects of childhood interventions on earnings), to marketing (e.g., the effects of incentives on long-term purchasing behavior). A common challenge in estimating such treatment effects is that long-term outcomes are typically either unobserved in the time frame needed to make policy decisions, or observed only for a small group of experimental subjects. One approach to overcome this missing data problem is to analyze treatments effects on an intermediate outcome, termed a "statistical surrogate" (Prentice 1989). The formal requirement for a variable to be a statistical surrogate, sometimes called the Prentice criterion (Begg and Leung 2000, Frangakis and Rubin 2002)), is independence of the treatment and the primary outcome conditional on the statistical surrogate. For example, in the case of studies of the effect of cancer therapies on mortality, tumor size serves as a statistical surrogate for mortality rates if mortality rates are independent of the treatment conditional on the value of the blood marker. Under this assumption, the treatment effect on mortality rates can be identified by from the relation between the treatment and tumor size and the relation between the tumor size and mortality rates using from a separate data set.

Although the use of surrogates has become widespread, the validity of the surrogacy condition is often controversial. Freedman et al. (1992) argued that the surrogate may not mediate all the effect of the treatment and developed a measure of the proportion of the treatment effect on the long-term outcome explained by the surrogate. Others have noted that unmeasured confounding between the surrogate and long-term outcome would invalidate the statistical surrogacy assumption, even if the treatment had no direct effect on the long-term outcome (Rosenbaum 1984, Frangakis and Rubin 2002, Joffe and Greene 2009, VanderWeele 2015).

In this paper, we approach this debate from a different perspective. Rather than attempting to determine whether the surrogacy condition holds for a given single intermediate outcome, we exploit that fact that in modern datasets, constructed from large scale electronic databases, researchers often observe a large number, possibly hundreds or thousands, of intermediate outcomes thought to lie on or close to the causal chain between the treatment and the long-term outcome of interest. These intermediate outcomes might be thought of as proxies for an unobserved latent true statistical surrogate. It may be that no individual candidate surrogate

satisfies the Prentice surrogacy criterion by itself, but that collectively these variables do satisfy the statistical surrogacy condition.

We focus primarily on 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 of interest, 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. Note that, in contrast to the study of mediation in causal problems, or the study of principal stratification, the surrogates are not of intrinsic interest in our analysis: their role is solely to aid in the identification and estimation of the average treatment effect of the treatment on the primary outcome.

As an example, consider evaluating the effects of early-childhood educational interventions, such as reductions in class size or improvements in teacher quality, on long-term outcomes, such as college attendance or earnings. Chetty et al (2011) estimated the effect of class size on earnings by linking data from the Tennessee Project STAR experiment, which randomized class size in grades kindergarten to third grade in the 1980s, to information on earnings decades later. The goal of our paper is to develop methods that will enable researchers to draw similar conclusions from educational experiments without waiting decades to observe the long- term outcome. In our framework, the experimental sample in this application would include data about class size (the treatment), student characteristics, and various intermediate outcomes (surrogates/proxies). The surrogates could include a variety of student outcomes in a few years following the treatment (e.g. grades and test scores across subject areas, as well as attendance). The observational sample would be a large panel dataset that would include the same student characteristics and surrogates as well as longer-term outcomes such as earnings.[1]

We consider three questions in this setting. First, how can the average treatment effect (ATE) be identified and estimated with a high-dimensional vector of surrogates that collectively satisfy the surrogacy assumption? Second, what is the bias from violations of the surrogacy assumption? Third, if the primary outcome is also observed in the experimental sample, is

[1]In another example, an internet company may be interested in the causal effect of a change in the user experience on long term engagement with the 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 page.

there still information to be gained from using surrogates?

To answer the first question, we introduce two new statistical concepts: the surrogate score, the probability of having received the treatment conditional on covariates and surrogates, and the surrogate index, defined as the expectation of the outcome of interest conditional on the surrogates. Under linearity, the surrogate index is a weighted average of each of the intermediate outcomes, with the weights determined by their ability to predict the primary outcome in the observational sample. We show that the ATE on the primary outcome can be identified by estimating the effect of the treatment on the surrogate index in the experimental sample under a set of assumptions. The key assumption is that the long-term outcome is independent of the treatment, conditional on the surrogates. In the class size application discussed above, the key requirement for identification of the ATE using surrogates is that (i) the test scores of the students in early grades capture all of the effects of the class size intervention and (ii) there are no unobserved confounders that affect both test scores and earnings. The ATE can also be estimated by averaging the outcomes in the observational sample using weights that depend on the surrogate score. Thus the surrogate index and surrogate score provide a simple way to collapse a high-dimensional vector of intermediate outcomes into a single index that can be used to estimate treatment effects, analogous to propensity scores (Rosenbaum and Rubin, 1983) in the causal inference literature. Also analogous to the propensity score literature, where different estimation methods may work better under different circumstances, whether methods based on the surrogate index or based on the surrogate score methods perform better depends on the empirical setting.

Next, we evaluate the degree of bias from the use of surrogates when the surrogacy condition fails. In this case, we show that our approach estimates an average causal effect on a function of the surrogate outcomes, where the function is the conditional expectation of the primary outcome given the surrogate outcomes in the observational sample. We then characterize the difference between this functional and the average treatment effect on the primary outcome itself. This characterization provides a method of assessing the potential degree of bias from violations of the surrogacy condition under alternative assumptions about how the treatment affects the primary outcome conditional on the intermediate outcomes. The formula for bias demonstrates why using many intermediate outcomes generally reduces the degree of bias. Intuitively, the degree of bias is determined by the extent to which the intermediates span the causal pathways

from the treatment to the primary outcomes. With a large and diverse set of intermediates, one is more likely to span all, or at least most of, these causal pathways. In the class size application, bias is likely to be smaller if there are many measures of student outcomes in the early grades, as well as a wide range of student characteristics that capture confounders that affect both surrogate outcomes and long-term outcomes. For example, the mapping from test scores to earnings may depend on parent income, in which case controlling for parent income would be valuable. In the limiting case where the intermediate outcomes perfectly predict either the primary outcome or the treatment, the bias vanishes.

Finally, we consider the case where the researcher observes the primary outcome in the experimental sample itself so that one can directly identify the average treatment effect on the primary outcome without making use of surrogates. However, there remains information content in the surrogates: using the surrogate index, one can estimate the average effect of interest generally more precisely. Building on the literature on semi-parametric estimation (e.g., Bickel, Klaassen, Ritov and Wellner, 1993), we establish the efficiency gain from the use of the surrogate index. The efficiency results show the conditions under which surrogates are most valuable for inference. They also clarify, for the two-sample case, how costly the lack of observations on the primary outcome in the experimental sample is. The use of surrogate indices is likely to be most useful in applications where the final outcome is a rare event or where substantial noise is introduced after intermediate outcomes are measured. In such settings – which include medical trials as well as experimentation (A/B testing) in other fields – using surrogate indices constructed from a battery of intermediate outcome can yield substantial gains by increasing precision.[2]

2 Set Up

As discussed in the introduction, this paper analyzes two distinct designs (single-sample and two-sample). In both cases the surrogacy assumption is valuable, although in different ways.

[2]As an example, Athey and Stern (2002) study the impact of Enhanced 911 adoption on cardiac patient outcomes, including mortality. Their data included a suite of surrogate patient health outcomes measured in the ambulance in addition to data about hospital outcomes including mortality (which occurred for only 3.5% of patients). They constructed a "health index" by projecting mortality on the surrogate health measures. Using the health index as a dependent variable rather than directly using mortality yielded gains in precision. Our efficiency results provide a formal justification for their approach and findings.

2.1 The Two Sample Design

Here we consider a setting with two samples, which we refer to as the two sample design (TSD). Motivated by the examples discussed in the Introduction, we refer to the first sample as the experimental sample and the second one as the observational sample. However, these are just labels, and we will make explicit any assumptions we make regarding the assignment and sampling in both samples.

The experimental and observational sample contain observations on N_E and N_O units, respectively. At times it will be convenient to view the data as consisting of a single sample of size $N = N_E + N_O$, with $P_i \in \{O, E\}$ a binary indicator for the group that unit i belongs to. For the N_E individuals in the experimental group there is a single binary treatment of interest $W_{E,i} \in \{0, 1\}$ and we are interested in the treatment's effect on a primary, often long-term, outcome, denoted by $Y_{E,i}$. To be precise in this two sample setting we index these variables by the sample, E or O, to which they belong. The outcome $Y_{E,i}$ is not observed. However, we do measure intermediate outcomes, which we refer to as surrogates (to be defined precisely in Section 3.2), denoted as $S_{E,i}$. Typically, the surrogate outcomes are vector-valued, and often the number of components will be substantial, in order to make the properties we propose feasible. Finally, we measure pre-treatment covariates $X_{E,i}$ for each individual. These variables are known not to be affected by the treatment.

Following the potential outcomes framework or Rubin Causal Model set up (Rubin, 2006, Holland, 1986; Imbens and Rubin, 2015), individuals in this group have two pairs of potential outcomes $(Y_{E,i}(0), Y_{E,i}(1))$ and $(S_{E,i}(0), S_{E,i}(1))$. We are interested in the causal effects on the outcome, $Y_{E,i}(1) - Y_{E,i}(0)$, typically an average of this over the population of interest. The realized outcomes are related to their respective potential outcomes as follows.

$$Y_{E,i} = \begin{cases} Y_{E,i}(0) & \text{if } W_{E,i} = 0, \\ Y_{E,i}(1) & \text{if } W_{E,i} = 1, \end{cases} \quad \text{and} \quad S_{E,i} = \begin{cases} S_{E,i}(0) & \text{if } W_{E,i} = 0, \\ S_{E,i}(1) & \text{if } W_{E,i} = 1. \end{cases}$$

Overall, all the units in population that the first sample is drawn from are characterized by the values of the sixtuple $(Y_{E,i}(0), Y_{E,i}(1), S_{E,i}(0), S_{E,i}(1), X_{E,i}, W_{E,i})$. For units in this sample we do not observe the full sixtuple. Rather, we observe only the triple $(S_{E,i}, X_{E,i}, W_{E,i})$ with support $\mathbb{S}_E$, $\mathbb{X}_E$, and $\mathbb{W}_E = \{0, 1\}$ respectively.

In the observational sample we do not know which treatment the N_O individuals were exposed to, and in fact, they need not be exposed to either treatment. For example, suppose we are

interested in the average causal effect of surgery versus a drug on a particular medical condition, with the experimental sample consisting of individuals exposed to either of those treatments. The observational sample may consist of individuals who neither took the drug, nor were exposed to surgery, possibly because the sample consists of observations from a time period when neither treatment existed. We observe a pretreatment variable $X_{O,i}$, the surrogate outcome $S_{O,i}$ and the primary outcome, $Y_{O,i}$, with support $\mathbb{Y}_O$, $\mathbb{S}_O$, and $\mathbb{X}_O$ respectively. We denote these variables in this sample using different labels from those for the corresponding variables in the experimental group because formally they need not measure the exact same object.

This set up with two samples, where the sets of variables that are observed in the two samples differs is implicit in much of the surrogacy literature. It is explicit in some studies on combining data sets, e.g., Ridder and Moffitt (2007) and Chen, Hong, and Tarozzi (2008). Rassler (2002,2004) refers to it as a data fusion setting. Graham, Campos de Xavier Pinto, and Egel (2016) discuss efficient estimation for a particular set of models defined by moment conditions in such a setting, where they allow $W_{E,i}$ to be a general random variable, rather than a binary indicator as in our set up.

2.2 The Single Sample Design

In the second setup we consider, there is a single population that is identical to the first population in the two-sample setup. All units in the population are characterized by the sextuple $(Y_{E,i}(0), Y_{E,i}(1), S_{E,i}(0), S_{E,i}(1), X_{E,i}, W_{E,i})$. For units in the sample we observe the quadruple $(S_{E,i}, X_{E,i}, W_{E,i}, Y_{E,i})$, now including the realized outcome $Y_{E,i}$. We refer to this setup as the single sample design (SSD).

Under the unconfoundedness assumption we discuss below, it is well known that the ATE is identified without further assumptions, and so statistical surrogacy does not play a role in identification. Nevertheless, the assumption can play an important role because it can make estimation and inference more precise.

2.3 The Estimand

We are interested in the average effect of the treatment on the outcome in the experimental group.

$$\tau = \mathbb{E}_E[Y_{E,i}(1) - Y_{E,i}(0)],$$

where to be explicit we index the expectation by the population the expectation is taken over. The fundamental problem for estimating τ in the experimental group is that the outcomes $Y_{E,i}$ are missing for all units in the experimental sample. We need to exploit the observational sample and its link to the experimental sample through the presence of the surrogate outcomes $S_{E,i}$. The surrogates, like the pretreatment variables, are not of intrinsic interest, and $\tau_S = \mathbb{E}_E[S_{E,i}(1) - S_{E,i}(0)]$ is of interest only in so far that it aids in estimation of τ.

3 Surrogacy and the Surrogate Score

In this section we discuss the surrogacy assumption and related concepts. To maintain the flow of the section we focus primarily on the two sample setting. The corresponding assumptions for the single sample setting are in most cases immediately clear. Whenever there are additional subtleties, we will point them out explicitly.

3.1 The Propensity Score and Unconfoundedness

Before we introduce the surrogacy assumption, we define some common quantities and assumptions in causal inference in observational studies (e.g., Rosenbaum, 2000; Imbens and Rubin, 2015). Specifically, for the individuals in the experimental group, we define the propensity score as the conditional probability of receiving the treatment (Rosenbaum and Rubin, 1983): $e(x) = \text{pr}_E(W_{E,i} = 1 | X_{E,i} = x)$. An assumption that is often invoked in observational studies is that the treatment assignment is unconfounded or ignorable conditional on the pre-treatment covariates and that there is overlap. Specifically, for individuals in the experimental group, we have:

Assumption 1. (Ignorable Treatment Assignment, Rosenbaum and Rubin, 1983) *(i)*

$$W_{E,i} \perp\!\!\!\perp \Big(Y_{E,i}(0), Y_{E,i}(1), S_{E,i}(0), S_{E,i}(1)\Big) \;\Big|\; X_{E,i},$$

(*ii*)

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

This assumption implies that in the experimental group, we could estimate the average causal effect of the treatment on the outcome $Y_{E,i}$ by adjusting for pretreatment variables, if the $Y_{E,i}$ were measured. There are many methods for implementing this. The original Rosenbaum and Rubin (1983) paper suggests matching or subclassification on the propensity score. Abadie and Imbens (2006) derive asymptotic properties for matching estimators. Hirano, Imbens and Ridder (2003) show that Horvitz-Thompson weighting estimators are efficient. Robins, Rotnitzky and Zhao (1995) develop what they call doubly robust estimators. See Rosenbaum (1995, 2002), Rubin (2006), Morgan and Winship (2007), and Imbens and Rubin (2015), for textbook discussions and reviews of this literature.

3.2 Statistical Surrogacy

Because the primary outcome $Y_{E,i}$ is not measured in the experimental group, we need to exploit the presence of the surrogates. The defining property of these surrogates $S_{E,i}$ is what Begg and Leung (2000) call the Prentice criterion, and what Frangakis and Rubin (2002) call statistical surrogacy, and which we simply refer to as surrogacy:

Assumption 2. (Surrogacy)

$$W_{E,i} \perp\!\!\!\perp Y_{E,i} \,\Big|\, S_{E,i}, X_{E,i}.$$

The literature following Prentice (1989) has been concerned with the plausibility of statistical surrogacy assumption and its relation to mediation (VanderWheele, 2015; Van Der Laan and Pedersen, 2004). Freedman et al. (1992) argued that the surrogate may not mediate all the effect of the treatment and provided a quantity to measure the proportion of effect on $Y_{E,i}$ explained by the surrogate $S_{E,i}$. Also, many noted that unmeasured confounding between $S_{E,i}$ and $Y_{E,i}$ and not captured by $X_{E,i}$ would invalidate the statistical surrogacy assumption, even if $W_{E,i}$ had no direct effect on $Y_{E,i}$ (Rosenbaum 1984, Frangakis and Rubin 2002, Joffe and Greene (2009), VanderWeele (2015)). Frangakis and Rubin (2002) developed a concept they labelled principal stratification to address questions related to mediation and surrogacy. Their starting point is

a candidate surrogate variable that is of substantive interest, in contrast to our setting where the surrogate is simply a means to an end. They develop a framework where adjusting for this candidate surrogate variable leads to causal effects of the treatment on the primary outcome. These are questions more closely aligned with those addressed in the mediation literature. See also Mealli and Mattei (2012) and Ding and Lu (2015).

We take a somewhat different perspective on the question of the validity of the surrogacy assumption. We view it as similar in spirit to the unconfoundedness assumption. It is unlikely to be satisfied exactly in any particular application, but, especially in cases with a large number of intermediate variables as well as pretreatment variables, it may be a reasonable approximation, as we will formalize in Section 4.2. Moreover, there is often no reasonable alternative. From our perspective it is useful to view the problem of identifying and estimating $\tau = \mathbb{E}_E[Y_{E,i}(1) - Y_{E,i}(0)]$ as a missing data one. The outcome $Y_{E,i}$ is missing for all units in the experimental sample, and any estimator of the treatment effect τ ultimately relies on imputing these missing outcomes. As we will formalize in Section 3.4, the surrogacy assumption is in that missing data perspective in essence an untestable missing-at-random assumption, conditional on the surrogates and the pretreatment variables. Any alternative assumption that is sufficiently strong to identify the average treatment effect must therefore violate the missing-at-random assumption even though there i no evidence against that assumption.

To exploit the notion of statistical surrogacy in settings with possibly many surrogates, we introduce a new concept, which we label the "surrogate score." It is the conditional probability of having received the treatment given the value for the surrogate outcome and the covariates.

Definition 1. (Surrogate Score)

$$r(s,x) = \text{pr}_E(W_{E,i} = 1 | S_{E,i} = s, X_{E,i} = x).$$

In contrast to the definition of the propensity score we write here 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. An important property the surrogate score shares with the propensity score is that it allows for statistical procedures that adjust only for scalar differences in other variables, irrespective of the dimension of the statistical surrogates. We state the next result without proof.

Proposition 1. (Surrogacy Score) *Under surrogacy (Assumption 2) we have*

$$W_{E,i} \perp\!\!\!\perp Y_{E,i} \,\Big|\, r(S_{E,i}, X_{E,i}).$$

3.3 Comparability of The Two Samples

This section discusses how we can use the information from the observational sample to help us estimate τ, specifically how to infer the missing values $Y_{E,i}$ in the experimental sample from the observed values $Y_{O,i}$ in the observational sample. Surrogacy is not sufficient for that, because that in itself does not make any assumptions about the observational sample. The key assumption is the conditional distribution of $Y_{E,i}$ given $(S_{E,i}, X_{E,i})$ is the same as the conditional distribution of $Y_{O,i}$ given $(S_{O,i}, X_{O,i})$. Formally,

Assumption 3. (Comparability of Samples)

$$Y_{E,i} \,\Big|\, S_{E,i}, X_{E,i} \;\sim\; Y_{O,i} \,\Big|\, S_{O,i}, X_{O,i},$$

and $\mathbb{X}_E = \mathbb{X}_O$, *and* $\mathbb{S}_E = \mathbb{S}_O$.

There are two immediate consequences of making the comparability assumptions, both of which allows us to share information between the two groups. To discuss these, we define the surrogate index:

Definition 2. (The Surrogate Index) *The surrogate index is the conditional expectation of the outcome given the surrogate outcomes and the pretreatment variables in the observational sample:*

$$h_O(s,x) = \mathbb{E}_O\left[Y_{O,i} \middle| S_{O,i} = s, X_{O,i} = x\right].$$

We can define the corresponding conditional expectation in the experimental sample:

$$h_E(s,x) = \mathbb{E}_E\left[Y_{E,i} \middle| S_{E,i} = s, X_{E,i} = x\right].$$

In contrast to $h_O(\cdot,\cdot)$, $h_E(\cdot,\cdot)$ is not estimable because we do not observe the outcome in the experimental sample. These conditional means 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 pretreatment variables. Define also the conditional expectation given treatment, pre-treatment variables and the surrogate:

$$\mu_E(s,x,w) = \mathbb{E}_E\left[Y_{E,i} \middle| S_{E,i} = s, X_{E,i} = x, W_{E,i} = w\right]. \tag{3.1}$$

We state the next result without proof.

Proposition 2. (Surrogate Index) (*i*) *Under surrogacy (Assumption 2) we have*

$$\mu_E(s,x,0) = \mu_E(s,x,1) = h_E(s,x), \qquad \text{for all } s \in \mathbb{S}_E,\ x \in \mathbb{X}_E.$$

(*ii*) *Under comparability (Assumption 3) we have*

$$\mathbb{S}_E = \mathbb{S}_O, \quad \mathbb{X}_E = \mathbb{X}_O, \quad \text{and } h_E(s,x) = h_O(s,x) \quad \text{for } s \in \mathbb{S}_E, \quad \text{and } x \in \mathbb{X}_E.$$

Next, let $q = N_E/(N_E + N_O)$ be the sampling weight of being in the experimental sample and $(1-q)$ be the sampling weight of being in the observational sample. Suppose we define the propensity to be in the experimental sample $P_i = E$ as follows

Definition 3. (Sampling Score)

$$t(s,x) = \frac{\text{pr}_E(S_{E,i} = s, X_{E,i} = x)q}{\text{pr}_E(S_{E,i} = s, X_{E,i} = x)q + \text{pr}_O(S_{O,i} = s, X_{O,i} = x)(1-q)}.$$

We also make the assumption

Assumption 4. Overlap in Sampling Score

$$t(s,x) < 1 \quad \text{for all } s, x$$

We can also also write $t(s,x) = \text{pr}(P_i = E | S_i = s, X_i = x)$, with a slight abuse of notation in defining a probability measure over P_i, which in our two sample design is not stochastic.

3.4 A Missing Data Approach

To get an intuition for the surrogacy and comparability assumptions, one can also frame them as a missing data assumption, close to the missingness at random (MAR) assumption common in the missing data literature (Rubin, 1976; Little and Rubin, 1988), and specifically the literature on combining samples with different sets of variables, (Gelman, King and Liu, 1998; Rassler,

2002; Rassler 2004; Graham, Campos de Xavier Pinto, and Egel, 2012, 2016). To see this, let $P_i = O$ indicating that the outcome was measured and $P_i = E$ otherwise, and define

$$Y_i = \begin{cases} Y_{E,i} & \text{if } P_i = E, \\ Y_{O,i} & \text{if } P_i = O, \end{cases} \qquad W_i = \begin{cases} W_{E,i} & \text{if } P_i = E, \\ W_{O,i} & \text{if } P_i = O, \end{cases}$$

$$S_i = \begin{cases} S_{E,i} & \text{if } P_i = E, \\ S_{O,i} & \text{if } P_i = O, \end{cases} \qquad X_i = \begin{cases} X_{E,i} & \text{if } P_i = E, \\ X_{O,i} & \text{if } P_i = O, \end{cases}$$

The complete data are $(X_i, S_i, Y_i, W_i, P_i)$. We view the sample as randomly drawn from a large population, so that we can view P_i as stochastic. For the units in the sample we observe the incomplete data $(X_i, S_i, 1_{P_i=O} \cdot Y_i, 1_{P_i=E} \cdot W_i, P_i)$. We can now rephrase the critical assumptions.

Assumption 5. (Missing Data Assumption)
Conditional on (S_i, X_i), the three variables P_i, Y_i and W_i are jointly independent, or, with some abuse of the Dawid conditional independence notation,

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

We state the following result without proof.

Proposition 3. (Missing Data Model)
(i) Assumption 5 implies Assumption 2 and 3,
and (ii) Assumption 5 has no testable implications.

Comparability corresponds to Y_i being independent of P_i given (S_i, X_i), and surrogacy corresponds to W_i being independent of Y_i given (S_i, X_i) and given $P_i = E$. Assumption 5 is in fact stronger than the combination of these two, because it also assumes that conditional on $P_i = O$, W_i is independent of Y_i, and it assumes that W_i is independent of P_i. Neither are required for our main results, but because we do not need the W_i in the observational sample and because these restrictions do not imply testable restrictions there is no loss of generality.

4 The Two Sample Design: Identification

4.1 Identification

Here we present two representations of the average treatment effect τ that suggest two different estimation strategies. Just as in the unconfoundedness setting the corresponding estimation

strategies differ in terms of the conditional expectations that need to be estimated. The full set of conditional expectations include the propensity score $e(x) = \text{pr}_E(W_{E,i} = 1|X_{E,i} = x)$, the surrogate score $r(s,x) = \text{pr}_E(W_{E,i} = 1|S_{E,i} = s, X_{E,i} = x)$, the sampling score $t(s,x) = \text{pr}(P_i = 1|S_i = s, X_i = x)$, and the surrogate index $h_O(s,x) = \mathbb{E}_O[Y_{O,i}|S_{O,i} = s, X_{O,i} = x]$.

The motivation for developing the different representations is that estimators corresponding to those different representations may have substantially different properties. Just as in the case of estimating average treatment effects under unconfoundedness, the lack of smoothness in the various scores or conditional expectations may affect the properties of estimators that rely on estimating these conditional expectations.

Define

$$\tau^E = \mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{W_{E,i}}{e(X_{E,i})} - h_O(S_{E,i}, X_{E,i}) \cdot \frac{1 - W_{E,i}}{1 - e(X_{E,i})}\right], \tag{4.1}$$

and

$$\begin{aligned}\tau^O = \mathbb{E}_O\Bigg[Y_{O,i} \cdot \frac{r(S_{O,i}, X_{O,i}) \cdot t(S_{O,i}, X_{O,i}) \cdot (1-q)}{e(X_{O,i}) \cdot (1 - t(S_{O,i}, X_{O,i})) \cdot q} \\ - Y_{O,i} \cdot \frac{(1 - r(S_{O,i}, X_{O,i})) \cdot t(S_{O,i}, X_{O,i}) \cdot (1-q)}{(1 - e(X_{O,i})) \cdot (1 - t(S_{O,i}, X_{O,i})) \cdot q}\Bigg],\end{aligned} \tag{4.2}$$

where the superscript on the τ indicates the population the expectation is taken over.

Theorem 1. *Suppose Assumptions 1, 2, 3, and 4 hold. Then,*

$$\tau \equiv \mathbb{E}_E[Y_{E,i}(1) - Y_{E,i}(0)] = \tau^E = \tau^O.$$

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. This will lead to an estimation strategy where in the experimental sample the missing $Y_{E,i}$ are imputed by $\hat{h}(S_{E,i}, X_{E,i})$. In contrast, 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, with the weights a function of the surrogate score and the sampling score. This will lead to an estimation strategy where in the observational sample the $Y_{O,i}$ are weighted proportional to the estimated surrogate score to estimate $\mathbb{E}_E[Y_{E,i}(1)]$, and weighted proportional to one minus the estimated surrogate

score to estimate $\mathbb{E}_E[Y_{E,i}(0)]$. There are additional representations, for example replacing $W_{E,i}$ in (4.1) by $r(S_{E,i}, X_{E,i})$, or replacing $Y_{O,i}$ in (4.1) by $h_O(S_{O,i}, X_{O,i})$. Estimators based on those representations do not appear to have attractive properties, either in theory or in our simulations.

4.2 The Consequences of Violations of Surrogacy and Comparability

In most applications the surrogacy assumption is at best a reasonable approximation. Instead the researcher may be confident that the assocation between the primary outcome and the treatment conditional on the proposed surrogate variables is limited, or just that there is a substantial association between the the surrogates and the primary outcome. In this section we interpret the probability limit of estimators based on either of the two characterizations of the estimand in Theorem 1 in case either or both of the surrogacy and comparability assumptions are violated. Throughout the section we maintain unconfoundedness.

Without surrogacy and comparability there are two things we can say.

Theorem 2. *First, (i)*

$$\tau^O = \tau^E = \mathbb{E}_E\left[h_O(S_{E,i}(1), X_{E,i}) - h_O(S_{E,i}(0), X_{E,i})\right],$$

and (ii), under unconfoundedness we have

$$\begin{aligned}
&\tau - \mathbb{E}_E\left[h_O(S_{E,i}(1), X_{E,i}) - h_O(S_{E,i}(0), X_{E,i})\right] \\
&\quad = \mathbb{E}\left[\Big\{\mu_E(S_{E,i}, X_{E,i}, 1) - \mu_E(S_{E,i}, X_{E,i}, 0)\Big\} \cdot \frac{r(S_{E,i}, X_{E,i}) \cdot (1 - r(S_{E,i}, X_{E,i}))}{e(X_{E,i}) \cdot (1 - e(X_{E,i}))}\right] \\
&\qquad + \mathbb{E}\left[\Big\{h_E(S_{E,i}, X_{E,i}) - h_O(S_{E,i}, X_{E,i})\Big\} \cdot \frac{r(S_{E,i}, X_{E,i}) - e(X_{E,i})}{e(X_{E,i}) \cdot (1 - e(X_{E,i}))}\right].
\end{aligned}$$

The first term captures the bias arising from violations of surrogacy, and the second term captures the bias arising from violations of comparability.

The first result shows that in general we estimate a valid causal effect as long as unconfoundedness holds. It is the average effect on a function of the surrogate, rather than the average effect on the primary outcome. This result also shows that which strategy we follow, using the surrogate score or the surrogate index to build an estimator, does not matter for the interpretation. The second result shows how lack of surrogacy and lack of comparability affect the

difference between what is being estimated and the average treatment effect on the outcome of interest.

Consider the bias from violations of surrogacy, the first term in the bias. It consists of two factors. The first factor is small if the surrogates explain much of the variation in $Y_{O,i}$ and therefore $\mu_E(s,x,1)$ and $\mu_E(s,x,0)$ are close. The second factor is small if the surrogate explains much of the variation in $W_{E,i}$, so that the surrogate score is close to zero or one and therefore $\mathbb{E}[r(S_{E,i},X_{E,i})\cdot(1-r(S_{E,i},X_{E,i}))]$ is close to zero.

Let us consider a special case where the assignment is completely random, so the propensity score is constant, $e(x)=p$, and where we have a substantial number of intermediate outcomes. These intermediate outcomes may be qualitatively very different, some continuous, some discrete or binary, and with very different substantive interpretations. The surrogate approach suggests a systematic way of combining the causal effects on the surrogates. Moreoever, suppose we approximate $h_O(s,x)$ by a linear function, $h_O(s,x)=\gamma_0+\gamma_S's+\gamma_X x$. Let $\tau_S=\mathbb{E}[S_{E,i}(1)-S_{E,i}(0)]$ be the average causal effect on the surrogates. Then τ^E can be estimated by

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

The linear model for $h_O(s,x)$ leads to a set of weights γ_S on the potentially large set of intermediate outcomes. Note the role of the pretreatment variables here. We do not simply regress the primary outcome on the surrogate outcomes. Instead we include the pretreatment variables in that regression, even if the data come from a randomized experiment, in order to improve the explanatory power of the surrogate index and the surrogate score.

It is also interesting to relate this discussion to the use of indices in health research. Consider the Body Mass Index (BMI), defined as (McGee et al, 2004; Adams et al, 2006). That index is defined as a person's weight in kilograms divided by their height in meters squared. This index is predictive of future health outcomes, although it is obviously not a conditional expectation. Nevertheless we can interpret estimates of the causal effect of treatments on the BMI through this approach.

5 The Two Sample Design: Estimation

In this section we discuss a number of estimation strategies. We take some of the insights from the literature on estimating average treatment effects under unconfoundedness to suggest

strategies that appear to be promising. The key difference with the unconfoundedness setting is that there are in the current setting two adjustments to be done.

5.1 An Estimator Based on the Surrogate Index

Suppose we estimate the surrogate index as $\hat{h}_O(s,x)$. We can then average this in the experimental sample for the treated and controls, after adjusting for the propensity score. A natural estimator, corresponding to (4.1), is the following difference of two average over the experimental sample:

$$\hat{\tau}^E = \frac{1}{\sum_{i=1}^{N_E} W_{E,i}/\hat{e}(X_{E,i})} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i},X_{E,i}) \cdot \frac{W_{E,i}}{\hat{e}(X_{E,i})} \tag{5.1}$$

$$- \frac{1}{\sum_{i=1}^{N_E}(1-W_{E,i})/(1-\hat{e}(X_{E,i}))} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i},X_{E,i}) \cdot \frac{1-W_{E,i}}{1-\hat{e}(X_{E,i})}.$$

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 the estimators substantially. In the case where the estimator for $h_O(s,x)$ was based on a linear specification, $h_O(s,x) = \gamma_0 + \gamma_S' s + \gamma_X' x$ is linear, this leads to

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

where $\hat{\tau}_S$ is an estimator for $\mathbb{E}_E\left[S_{E,i}(1) - S_{E,i}(0)\right]$. In the case without pretreatment variables where the experimental sample came from a completely randomized experiment, this would further simplify to

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

where $\overline{S}_1$ and $\overline{S}_0$ are the average values for the surrogate outcome in treated and control samples respectively. However, we emphasize that in general, there may be interactions between the surrogates and pre-treatment variables.

When the number of pre-treatment variables or surrogates (and their interactions) is large, using logistic regression may not be feasible, and one may wish to consider regularization methods such as LASSO (Tibshirani, 1996; Belloni, Chernozhukov and Hansen, 2014), ridge regression, tree or forest based methods (Breiman, Friedman, Olshen, and Stone, 1984; Wager and

Athey, 2015), or super learners (VanderLaan and Rose, 2011) to estimate the various scores and conditional expectations.

5.2 An Estimator Based on the Surrogate Score

In this Section we use the second representation for τ in the main theorem. 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. Note that we have assumed the most typical models for the propensity score, the surrogate score, and the sampling score and there is no doubt that our resulting estimate of the treatment effect could be sensitive to misspecification of these models especially if there is limited overlap. However, we feel this would provide a starting point for estimating the treatment effect under our setting. Again in settings with a large number of surrogates or pretreatment variables one may wish to use regularization methods. Once we have estimates $\hat{e}(x)$, $\hat{r}(s,x)$ and $\hat{t}(s,x)$, we would plug them into the sample analogs of the expected values in the main theorem.

What we refer to as the surrogate score estimator is based on averaging over the observational sample:

$$\hat{\tau}^O = \frac{1}{\sum_{i=1}^{N_O} \omega_{1,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_O} Y_{O,i} \cdot \omega_{1,\hat{r},\hat{e},\hat{t}} - \frac{1}{\sum_{i=1}^{N_O} \omega_{0,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_O} Y_{O,i} \cdot \omega_{0,\hat{r},\hat{e},\hat{t}}, \tag{5.2}$$

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

$$\omega_{w,\hat{r},\hat{e},\hat{t}} = \frac{\hat{r}(S_{O,i}, X_{O,i})^w \cdot (1-\hat{r}(S_{O,i}, X_{O,i}))^{1-w} \cdot \hat{t}(S_{O,i}, X_{O,i}) \cdot (1-q)}{\hat{e}(X_{O,i})^w \cdot (1-\hat{e}(X_{O,i}))^{1-w} \cdot (1-\hat{t}(S_{O,i}, X_{O,i})) \cdot q}.$$

5.3 Matching Estimators

Although matching estimators are generally not efficient in settings with unconfoundedness (Rubin, 2006; Abadie and Imbens, 2006, 2016), they have a lot of intuitive appeal, and it is instructive to see how a matching strategy could be implemented in this case. Consider unit i in the experimental sample with $X_{E,i} = x$ and $S_{E,i} = s$, and suppose this is a treated unit with $W_{E,i} = 1$. 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. Specfically, we need to find the closest unit in the experimental sample, in terms of pretreatment variables, among the units with $W_{E,i} = 0$. Suppose this unit is unit j, with $W_{E,j} = 0$, and the value of the pretreatment variable for this unit is $X_{E,j} = x'$, and the surrogate is $S_{E,j} = s'$ (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. First, find the unit in the observational sample closest to unit i, in terms of both pretreatment variables and surrogates. Let $i'(i)$ be the index for this unit, and let the value of the outcome for this unit be $Y_{O,i'}$, and the values of the pretreatment variables and surrogates $X_{O,i'}$ and $S_{O,i'}$ (now as a result of the matching $X_{O,i} \approx X_{O,i'}$ and $S_{O,i} \approx S_{O,i'}$. Finally, find the unit in the observational sample closest to unit j, in terms of both pretreatment variables and surrogates. Let the value of the outcome for this unit be $Y_{O,j'}$, and the values of the pretreatment variables and surrogates $X_{O,j'}$ and $S_{O,j'}$, with $X_{O,j} \approx X_{O,j'}$ and $S_{O,j} \approx S_{O,j'}$.

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

$$\widehat{Y_{E,i}(1) - Y_{E,i}(0)} = Y_{O,i'} - Y_{O,j'}.$$

The matching estimator for τ would then be the average of this over the experimental sample.

In settings with high-dimensional pre-treatment variables or surrogates this matching strategy it would be unlikely that such a matching strategy would be effective, and methods relying on regularized estimation of the surrogate index or surrogate score would be more attractive.

6 Simulation

6.1 Setup

We conduct a small simulation study to assess the performance of different estimation methods for τ if the identifying assumptions are met. To focus on the role of the surrogate variables, we constrain the study to a randomized experimental design without pre-treatment covariates so that the propensity score is constant, $e(x) = p$, and a constant sampling score so that $q = t(s, x)$. Within this design we focus on the role of the surrogate index and the surrogate score. Specifically, let $\hat{h}_O(\cdot)$ be the ordinary least squares estimate of the conditional expectation of $Y_{O,i}$ given $S_{O,i}$ and let $\hat{r}(\cdot)$ be the logistic regression estimate of the conditional expectation of $W_{O,i}$ given $S_{O,i}$. We study the following two estimators for τ, simplified versions of (5.1)-(5.2), to the case with $e(x) = p$ and $t(s, x) = q$:

$$\hat{\tau}^O = \frac{1}{\sum_{i=1}^{N_O} \hat{r}(S_{O,i})} \sum_{i=1}^{N_O} Y_{O,i} \cdot \hat{r}(S_{O,i}) - \frac{1}{\sum_{i=1}^{N_O} (1 - \hat{r}(S_{O,i}))} \sum_{i=1}^{N_O} Y_{O,i} \cdot (1 - \hat{r}(S_{O,i}))$$

$$\hat{\tau}^E = \frac{1}{\sum_{i=1}^{N_E} W_{E,i}} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i}) \cdot W_{E,i} - \frac{1}{\sum_{i=1}^{N_E} (1 - W_{E,i})} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i}) \cdot (1 - W_{E,i})$$

Subsequent sections study the behaviors of $\hat{\tau}^O$ and $\hat{\tau}^E$ under different data generating processes. In particular, we study (i) the properties of the surrogate score and the surrogate index as the number of surrogates increases, (ii) the consequences of misspecifying the surrogate score and the surrogate index, (iii) the role of different sample sizes in different samples, and (iv) the role of the explanatory power of the surrogates in the surrogacy score and the surrogacy index. In all simulation settings, we study the bias and variance of the two estimators $\hat{\tau}^O$ and $\hat{\tau}^E$ evaluated from 1000 simulated data sets.

6.2 Dimension of Surrogates

In this section, we consider the effect of increasing the dimension of the surrogates on estimating τ. Each data set has $N = 1000$ individuals with 500 from the experimental sample and 500from the observational sample. Suppose we have M surrogates, where M takes on values from 1 to 200. The M surrogates follow a multivariate standard Normal with mean zero and identity covariance under both the observational and the experimental sample. We generate data based

on the following model.

$$P(W_{E,i}=1|S_{E,i}) \sim \mathcal{B}\left(1, \frac{\exp(\alpha_0+\alpha_S' S_{E,i})}{1+\exp(\alpha_0+\alpha_S' S_{E,i})}\right), \quad P(Y_{O,i}|S_{O,i}) \sim \mathcal{B}\left(1, \frac{\exp(\gamma_0+\gamma_S' S_{E,i})}{1+\exp(\gamma_0+\gamma_S' S_{E,i})}\right)$$

where α_S are fixed parameters chosen from a standard Normal with mean 0 and variance $1/M$ and $\gamma_S = \alpha_S$. We also generate $Y_{E,i}$ under the same model as $Y_{O,i}$. For the experimental sample, we only use $(W_{E,i}, S_{E,i})$ and in the observational sample, we only use $(S_{O,i}, Y_{O,i})$. Note that all of the identifying assumptions are satisfied by the simulation design.

Figure 1 shows the result of the simulation. We see that regardless of the dimension of M,

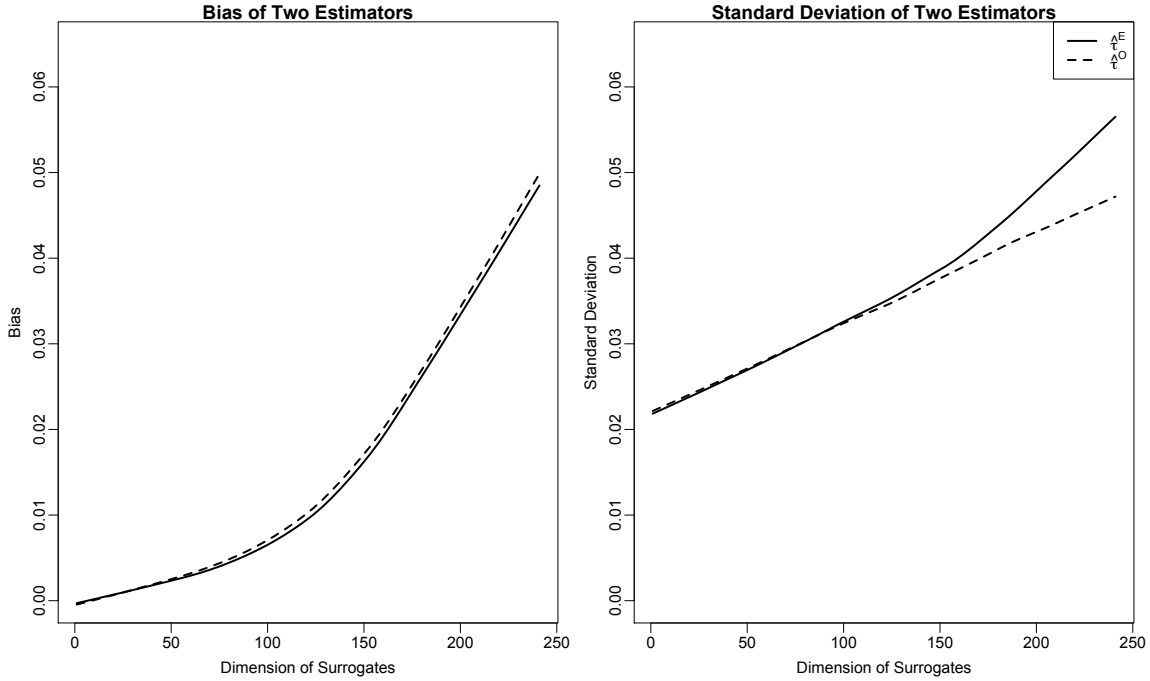

Figure 1: Simulation study of the dimension of surrogates. Bias is the absolute value of the empirical bias over $1,000$ simulations. Standard deviation is the empirical standard deviation of the estimates over $1,000$ simulations.

both estimators have similar performance with respect to bias and variance, although $\hat{\tau}^E$ has a slightly higher variance as dimension of the surrogates are quite large. Also, as expected, the bias and variance from both estimators increase as the dimension of the surrogates grows because the sample size remains fixed at $N_E = 500$ and $N_O = 500$. In short, the simulation demonstrates that the estimation methods can handle large number of surrogates at the expected loss in bias and variance.

6.3 Misspecification

In this section, we consider the effect of using an inadequate number of surrogates. In our set up there are 250 surrogates that collectively satisfy the surrogacy assumption. We then compare the two estimators, using only the first K surrogates, for $K = 1, \ldots, 250$. The sample size remains fixed at 1,000, with $N_E = 500$ and $N_O = 500$. The coefficients on the surrogate variables are $\alpha_{S,k} = \gamma_{S,k} = (1/3) \cdot k^{-1/2}$, so that the initial surrogates are the most important ones.

Figure 2 shows the result of the simulation.

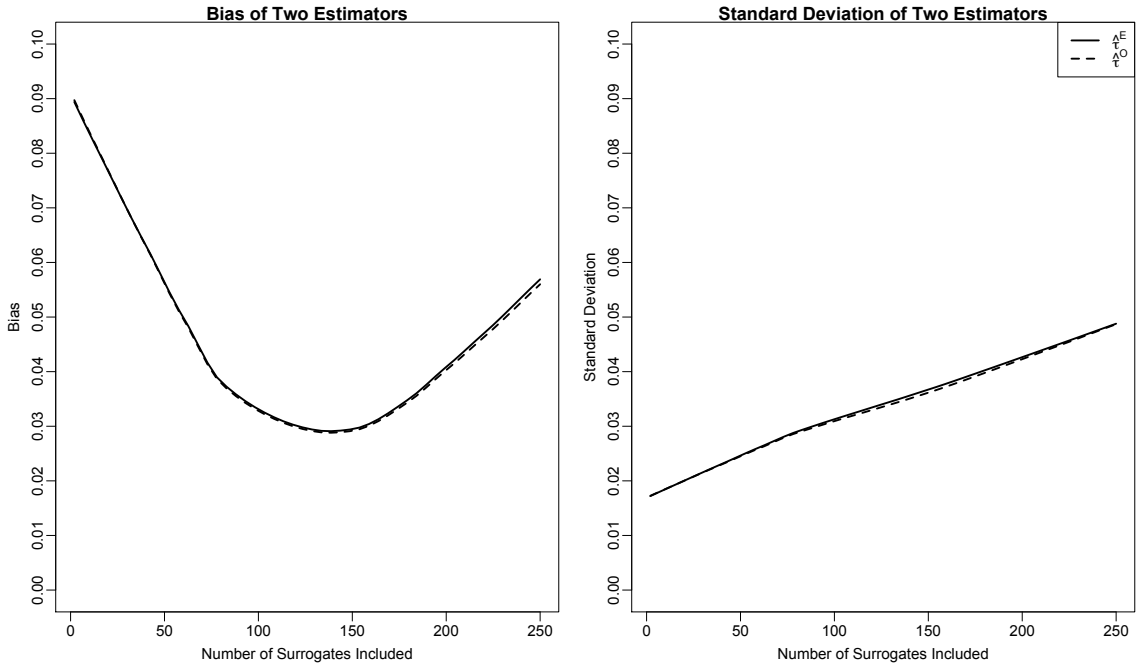

Figure 2: Simulation study of the effects of insufficient surrogates. Bias is the absolute value of the empirical bias over $1,000$ simulations. Standard deviation is the empirical standard deviation of the estimates over $1,000$ simulations.

We see that initially increasing the number of surrogates improves the bias of both estimators. As the number of surrogates increases, at some point the remaining surrogates contribute too little information to improve the bias, and small sample issues start dominating. At that point the bias starts increasing with the number of covariates, just as in the earlier simulations where the set of surrogates used was always sufficient.

6.4 Different Sample Sizes

In this section, we consider the effect of having different sample sizes from different samples in estimation. The simulation setup is identical to Section 6.2 except we fix $M = 10$, set α_S and γ_S so that $\alpha_S = \gamma_S$ and the treatment effect is equal to 0.5, and vary $q = N_E/(N_E + N_O)$, the relative proportion of the experimental sample. A $q < 1/2$ implies that there are more units in the observational data than the experimental data while a $q > 1/2$ implies that there are more units in the experimental data than the observational data. At $q = 1/2$, the sample sizes between the experimental and the observational samples are identical. We vary q from 0.05 to 0.95 and study the estimation properties of $\hat{\tau}^O$ and $\hat{\tau}^E$ under this setting.

	$\hat{\tau}^O$		$\hat{\tau}^E$	
q	Bias	Standard Deviation	Bias	Standard Deviation
0.05	2.011	6.357	0.023	7.490
0.25	0.001	3.018	0.060	3.508
0.5	0.013	2.801	0.012	2.850
0.75	0.067	3.482	0.012	3.004
0.95	0.423	7.420	2.747	6.434

Table 1: Simulation study of different sample sizes. Bias is the absolute value of the empirical bias over 1000 simulations. Standard deviation is the empirical standard deviation of the estimates across 1000 simulations. All values are multiplied by 100 for easy reading.

Table 1 summarizes the results. When the sample sizes are roughly equivalent in both the observational and the experimental sample, we achieve the lowest variance for both estimators and the variance for both estimators form bowl-shape as we vary q. However, bias fluctuates depending on q and the estimator. For example, bias is the highest for $\hat{\tau}^O$ when $q = 0.05$, perhaps because the surrogate score is poorly estimated due to the small sample size of the experimental data even though there is a lot of samples in the observational data. Similarly, the bias for $\hat{\tau}^E$ is the highest when $q = 0.95$, most likely because the surrogate index is poorly estimated from the small sample size of the observational data. However, for $\hat{\tau}^O$, even if $q = 0.95$ and we have a better estimate of the surrogate score, there is still more bias compared to $q = 0.75$ or $q = 0.5$ since there isn't enough samples in the observational. data. A similar phenomena can be observed with $\hat{\tau}^E$ when $q = 0.05$ and we have a good estimate of the surrogate index, although the bias of $\hat{\tau}^E$ at $q = 0.05$ is less pronounced than that of $\hat{\tau}^O$ at $q = 0.95$. Indeed, when

it comes to bias, the simulation suggests a complex non-linear trade-off between obtaining good estimates of the surrogate score/index and having enough samples in the other data to utilize these estimated scores/indices.

6.5 Explanatory Power

In this section, we characterize the behavior of the two estimators when we increase the explanatory power of the surrogate score and the index. The simulation setup is identical to Section 6.2 except we fix $M = 10$ and we set α_S and γ_S based on the following distributions laid out in Table 2.

	$\hat{\tau}^O$		$\hat{\tau}^E$	
Design	Bias	Standard Deviation	Bias	Standard Deviation
$\alpha_S \sim N(0, 1/M)$, $\gamma_S \sim N(0, 1/M)$	0.030	2.191	0.022	2.214
$\alpha_S \sim N(0, 4/M)$, $\gamma_S \sim N(0, 1/M)$	0.235	3.407	0.137	3.448
$\alpha_S \sim N(0, 1/M)$, $\gamma_S \sim N(0, 4/M)$	0.169	3.089	0.093	3.162
$\alpha_S \sim N(0, 4/M)$, $\gamma_S \sim N(0, 4/M)$	0.222	3.566	0.111	3.581

Table 2: Simulation study of explanatory power. Bias is the absolute value of the empirical bias over 1000 simulations. Standard deviation is the empirical standard deviation of the estimates across 1000 simulations. All values are multiplied by 100 for easy reading.

As expected, we see that as the variance of α_S and γ_S increase, the variance of both estimators increases, although obviously if the surrogates have very little explanatory power the variance must increase. The story for bias is a bit more complex. Bias tends to be the lowest when the variance of α_S and γ_S is small, with the exception of the estimator $\hat{\tau}^E$, which has lower bias than its counterpart $\hat{\tau}^O$. Note that the bias of $\hat{\tau}^E$ is affected by the variance increase in any one of the parameters γ_S and α_S.

6.6 Summary

In summary, the simulation study reveals the following trends. First, while fixing the sample size, if one increases the dimensions of the surrogates, $\hat{\tau}^O$ outperforms $\hat{\tau}^E$ in terms of variance. Second, the sensitivity to misspecification is similar. Third, when the sample sizes between the two data sets differ, there is a interesting trade-off between bias and variance for both estimators.

For example, variance tends to be minimized when there is an equal sample size between the two data sets and bias tends to be minimized at non-extreme, but not necessarily equal, sample sizes. The modelling assumptions, when correct, are more valuable for the smallest of two two samples, so that if the experimental sample is smaller than the observational sample, $\hat{\tau}^E$ outperforms $\hat{\tau}^O$. Fourth, the explanatory power simulation suggests that when α_S and γ_S are drawn from distributions with higher variance the bias tends to be small for $\hat{\tau}^E$ compared to $\hat{\tau}^O$. The simulation study, especially the one concerning unequal sample size, hints at the complexity of estimation and finite-sample performance of these estimators and we leave it as an area of future research to precisely characterize properties of estimators.

7 The Single Sample Design: Efficiency

In this section we consider the single sample design, and analyze the potential for efficiency gains that might arise by exploiting the surrogacy assumption. We use our findings to further quantify the efficiency losses that arise due to the failure to observe the long-term outcome in the two-sample setting. Focusing on the information content from the surrogacy assumption, our semiparametric efficiency bound analysis follows in the spirit of Bickel, Klaassen, Ritov and Wellner (1993).

7.1 Efficiency Bounds: The Value of Surrogacy

In the single sample case, in the absence of covariates and without further assumptions, it is well known that an efficient estimator for the effect of a treatment $W_{E,i}$ on $Y_{E,i}$ is the difference between the sample mean of the treated outcomes and the sample mean of the control outcomes. Thus, it might seem that incorporating surrogate variables $S_{E,i}$ in estimation (for example, by replacing $Y_{E,i}$ by the surrogate index in estimation, as in τ^E) would hurt efficiency. However, in this section we show that the opposite is true, once we incorporate the surrogacy assumption. The intuition is that the surrogacy assumption allows us to pool all data-including data for both treated and control units-when estimating the relationship between $S_{E,i}$ and $Y_{E,i}$, since the surrogacy assumption requires that this relationship does not vary with the treatment.

Let $\sigma^2(s,x) = \mathbb{V}_E(Y_{E,i}|S_{E,i} = s, X_{E,i} = x)$, $\sigma_w^2(x) = \mathbb{V}_E(Y_{E,i}|X_{E,i} = x, W_{E,i} = w)$, and $\mu_w(x) = \mathbb{E}_E[Y_{E,i}|X_{E,i} = x, W_{E,i} = w]$. Then, we have the following efficiency result.

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

$$\begin{aligned}\mathbb{V}_{ns} &= \mathbb{E}_E\left[\frac{\sigma_1(X_{E,i})^2}{e(X_{E,i})} + \frac{\sigma_0^2(X_{E,i})}{1-e(X_{E,i})} + \left(\mu_1(X_{E,i}) - \mu_0(X_{E,i}) - \tau\right)^2\right] \\ &= \mathbb{E}_E\Bigg[\sigma^2(S_{E,i}, X_{E,i}) \cdot \left(\frac{r(S_{E,i}, X_{E,i})}{(e(X_{E,i}))^2} + \frac{1-r(S_{E,i}, X_{E,i})}{(1-e(X_{E,i}))^2}\right) \\ &+ \frac{r(S_{E,i})}{(e(X_{E,i}))^2} \cdot \left(h_E(S_{E,i}, X_{E,i}) - \mu_1(X_{E,i})\right)^2 + \frac{1-r(S_{E,i}, X_{E,i})}{(1-e(X_{E,i}))^2} \cdot \left(h_E(S_{E,i}, X_{E,i}) - \mu_0(X_{E,i})\right)^2 \\ &\quad + \left(\mu_1(X_{E,i}) - \mu_0(X_{E,i}) - \tau\right)^2\Bigg].\end{aligned}$$

(ii) The efficiency bound assuming surrogacy is

$$\begin{aligned}\mathbb{V}_{s} &= \mathbb{E}_E\Bigg[\sigma^2(S_{E,i}, X_{E,i}) \cdot \left(\frac{r^2(S_{E,i}, X_{E,i})}{(e(X_{E,i}))^2} + \frac{(1-r(S_{E,i}, X_{E,i}))^2}{(1-e(X_{E,i}))^2}\right) \\ &+ \frac{r(S_{E,i})}{(e(X_{E,i}))^2} \cdot \left(h_E(S_{E,i}, X_{E,i}) - \mu_1(X_{E,i})\right)^2 + \frac{1-r(S_{E,i}, X_{E,i})}{(1-e(X_{E,i}))^2} \cdot \left(h_E(S_{E,i}, X_{E,i}) - \mu_0(X_{E,i})\right)^2 \\ &\quad + \left(\mu_1(X_{E,i}) - \mu_0(X_{E,i}) - \tau\right)^2\Bigg].\end{aligned}$$

The difference between the two bounds, $\mathbb{V}_n - \mathbb{V}_c$, is the efficiency gain from exploiting surrogacy. The expressions differ in the first term, involving $\sigma^2(S_{E,i}, X_{E,i})$. There is no gain if $S_{E,i} = W_{E,i}$ (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).

To gain more intuition about where the gain is biggest, we can write the difference in the efficiency bounds, assuming homoskedasticity so that $\sigma^2 = \sigma^2(s,x)$ for all s and x and no pretreatment variables, as

$$\mathbb{V}_n - \mathbb{V}_c = \mathbb{E}\left[\frac{2\cdot\sigma^2}{p\cdot(1-p)} \cdot \left\{p\cdot(1-p) - (r(S_i) - p)^2\right\}\right].$$

where $p = \mathbb{E}_E[W_{E,i}]$. Again, there is no gain if $S_{E,i} = W_{E,i}$ so that $r(S_{E,i}) \in \{0,1\}$, and the gain is biggest if $r(S_i)$ constant (and thus equal to $\mathbb{E}_E[r(S_{E,i})] = p$). Interestingly, recalling Theorem 2, when $r(S_{E,i})$ is close to 0 or 1, then the bias due to failure of the statistical surrogacy is small, while when $\mathbb{E}_E[r(S_{E,i})]$ is close to p, the bias due to the failure of comparability is small. Thus, for applications where $\mathbb{E}_E[r(S_{E,i})]$ is close to p and the statistical surrogacy assumption is very credible, then even if there are possible violations of comparability, using the surrogate index approach to estimation rather than directly estimating the effect of the treatment on the final outcome may improve efficiency without creating much bias.

7.2 Efficiency Bounds: The Value of Observing the Primary Outcome

In this section, we calculate the efficiency bound for the single sample design when for part of the sample Y_i is missing and for the remainder of the sample W_i is missing. For simplicity, we focus on the case without pretreatment variables, and assume that the sampling score is constant.

Theorem 4. *Suppose $t(s,x) = q$. Then in the Two Sample Design the efficiency bound is*

$$\mathbb{V}_s = \mathbb{E}\left[\frac{\sigma^2(S_i)}{1-q}\cdot\left(\frac{r(S_i)}{p^2}+\frac{1-r(S_i)}{(1-p)^2}-2\cdot\frac{r(S_i)\cdot(1-r(S_i))}{p^2\cdot(1-p)^2}\right)\right.$$
$$\left.+\frac{1}{q}\cdot\left\{\frac{r(S_i)}{p}\cdot(\mu(S_i)-\mu_1)^2+\frac{1-r(S_i)}{1-p}\cdot(\mu(S_i)-\mu_0)^2\right\}\right].$$

The first term in the efficiency bound in the Single Sample Design increases by a factor $1/(1-q)$, and the second factor increases by a factor $1/q$. Depending on the value of the two terms and the value of q the efficiency loss from not observing the outcome and the treatment in the same sample may be modest or very large. For example, if the sampling probability q is small and the variance of outcomes conditional on the surrogates and the treatment status is large, the loss from failing to observe outcomes is large.

8 Conclusion

In this paper we analyze the role of surrogates in estimating average treatment effects. We focus on two cases. In the first we have two samples, one where we observe the treatment and the surrogate variables, and one where we observe the surrogate variables and the outcome of interest. We formalize assumptions under which we can identify the average effect of the treatment on the outcome, thus providing guidance on how to select surrogates and how to reason about whether estimation approaches based on the surrogate index and the surrogate score would be justified. For cases where the assumptions may be controversial, we characterize the bias due to different types of violations of our assumptions, and in cases where the final outcome is bounded (e.g. when it is binary), we can bound the bias. We further propose estimation strategies that may be effective when there are many surrogates of pre-treatment

variables; the surrogate index or the surrogate score can be estimated using regularized regression or other high-dimensional estimation methods to allow for dimensionality reduction. We also consider the case where we observe all variables in a single sample, and derive the information gain from surrogacy assumptions. Our results imply that using the surrogate index approach may be more efficient than focusing on final outcomes, even in a single sample where the final outcomes are observed.

Appendix

Proof of Theorem 1
We write $\tau = \mathbb{E}_E[Y_{E,i}(1)] - \mathbb{E}_E[Y_{E,i}(0)]$. The results are implied by the following equalities,

$$\mathbb{E}_E[Y_{E,i}(1)] = \mathbb{E}_O\left[Y_{O,i} \cdot \frac{r(S_{O,i}, X_{O,i}) \cdot t(S_{O,i}, X_{O,i}) \cdot (1-q)}{e(X_{O,i}) \cdot (1 - t(S_{O,i}, X_{O,i})) \cdot q}\right], \tag{A.1}$$

$$\mathbb{E}_E[Y_{E,i}(1)] = \mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{W_{E,i}}{e(X_{E,i})}\right]. \tag{A.2}$$

We prove one of them, the others are similar, and proofs are available from the authors.
Consider (A.2). By Assumption 1 (ignorable treatment assignment), it follows that

$$\mathbb{E}_E[Y_{E,i}(1)] = \mathbb{E}_E\left[Y_{E,i} \cdot \frac{W_{E,i}}{e(X_{E,i})}\right].$$

Using the law of iterated expectations, we can first condition on $S_{E,i}$ and $X_{E,i}$ to get

$$\mathbb{E}_E\left[Y_{E,i} \cdot \frac{W_{E,i}}{e(X_{E,i})}\right] = \mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i} \cdot \frac{W_{E,i}}{e(X_{E,i})}\middle| S_{E,i}, X_{E,i}\right]\right].$$

By Assumption 2 (surrogacy), we have

$$\mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i} \cdot \frac{W_{E,i}}{e(X_{E,i})}\middle| S_{E,i}, X_{E,i}\right]\right] = \mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i}|S_{E,i}, X_{E,i}\right] \cdot \frac{\mathbb{E}_E\left[W_{E,i}|S_{E,i}, X_{E,i}\right]}{e(X_{E,i})}\right]$$

By Assumption 3 (comparabiliity), $h_O(s,x) = h_E(s,x)$ so that this is equal to

$$\mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{\mathbb{E}_E\left[W_{E,i}|S_{E,i}, X_{E,i}\right]}{e(X_{E,i})}\right]$$

Un-doing the law of iterated expectations gives us the desired equality. □

Proof for Theorem 2 We focus on part (ii). The proof for (i) is available from the authors.
By definition,

$$\tau = \mathbb{E}_E[Y_{E,i}(1) - Y_{E,i}(0)] = \mathbb{E}_E[Y_{E,i}(1)] - \mathbb{E}_E[Y_{E,i}(0)].$$

By unconfoundedness, this is equal to

$$\tau = \mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i}|W_{E,i}=1, X_{E,i}\right]\right] - \mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i}|W_{E,i}=0, X_{E,i}\right]\right].$$

By iterated expectations this is equal to

$$\begin{aligned}\tau = \mathbb{E}_E\left[\mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i}|S_{E,i},X_{E,i},W_{E,i}=1\right]|W_{E,i}=1,X_{E,i}\right]\right.\\ \left.-\mathbb{E}_E\left[\mathbb{E}_E\left[\mathbb{E}_E\left[Y_{E,i}|S_{E,i},X_{E,i},W_{E,i}=0\right]|W_{E,i}=0,X_{E,i}\right]\right.\right.\\ = \mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E\left(S_{E,i},X_{E,i},1\right)|W_{E,i}=1,X_{E,i}\right]\right]\\ -\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E\left(S_{E,i},X_{E,i},0\right)|W_{E,i}=0,X_{E,i}\right]\right].\end{aligned}$$

Thus, defining

$$\tau_m = \mathbb{E}_E\left[h_O(S_{E,i}(1),X_{E,i}) - h_O(S_{E,i}(0),X_{E,i})\right],$$

we have

$$\begin{aligned}\tau-\tau_m = \mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},1)|W_{E,i}=1,X_{E,i}\right]\right.\\ -\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},0)|W_{E,i}=0,X_{E,i}\right]\right.\\ \left.-\Big\{\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]\Big\}.\end{aligned}$$

Add and subtract

$$\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right],$$

to get

$$\begin{aligned}\tau-\tau_m = \mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},1)|W_{E,i}=1,X_{E,i}\right]\right.\\ -\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},0)|W_{E,i}=0,X_{E,i}\right]\right.\\ -\Big\{\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]\Big\}\\ +\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]\\ \left.-\Big\{\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]\Big\}.\right.\end{aligned}$$

Rearranging the terms this is equal to

$$\tau-\tau_m = \mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},1)|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right] \quad \text{(A.3)}$$

$$-\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i},X_{E,i},0)|W_{E,i}=0,X_{E,i}\right]\right]+\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right] \quad \text{(A.4)}$$

$$+\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=1,X_{E,i}\right]\right] \quad \text{(A.5)}$$

$$+\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]-\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i},X_{E,i})|W_{E,i}=0,X_{E,i}\right]\right]. \quad \text{(A.6)}$$

Next, note that by definition of expectations,

$$\begin{aligned}h_E(s,x) = \mathbb{E}[Y_{E,i}|S_{E,i}=s,X_{E,i}=x]\\ = \mathbb{E}[Y_{E,i}|S_{E,i}=s,X_{E,i}=x,W_{E,i}=1]\cdot\mathrm{pr}(W_{E,i}=1|S_{E,i}=s,X_{E,i}=x)\\ +\mathbb{E}[Y_{E,i}|S_{E,i}=s,X_{E,i}=x,W_{E,i}=0]\cdot\mathrm{pr}(W_{E,i}=0|S_{E,i}=s,X_{E,i}=x)\end{aligned}$$

$$= \mu_E(s,x,1) \cdot r(s,x) + \mu_E(s,x,0) \cdot (1 - r(s,x)).$$

Use this to write (A.3) as

$$\begin{aligned}
&\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i}, X_{E,i}, 1)|W_{E,i} = 1, X_{E,i}\right]\right] \\
&-\mathbb{E}_E\left[\mathbb{E}_E\left[\mu_E(S_{E,i}, X_{E,i}, 1) \cdot r(S_{E,i}, X_{E,i}) + \mu_E(S_{E,i}, X_{E,i}, 0) \cdot (1 - r(S_{E,i}, X_{E,i}))|W_{E,i} = 1, X_{E,i}\right]\right] \\
&= \mathbb{E}_E\left[\mathbb{E}_E\left[\left\{\mu_E(S_{E,i}, X_{E,i}, 1) - \mu_E(S_{E,i}, X_{E,i}, 0)\right\} \cdot (1 - r(S_{E,i}, X_{E,i}))|W_{E,i} = 1, X_{E,i}\right]\right] \\
&= \mathbb{E}_E\left[\mathbb{E}_E\left[\left\{\mu_E(S_{E,i}, X_{E,i}, 1) - \mu_E(S_{E,i}, X_{E,i}, 0)\right\} \cdot \frac{(1 - r(S_{E,i}, X_{E,i})) \cdot r(S_{E,i}, X_{E,i})}{e(X_{E,i})}\middle| X_{E,i}\right]\right] \\
&= \mathbb{E}_E\left[\left\{\mu_E(S_{E,i}, X_{E,i}, 1) - \mu_E(S_{E,i}, X_{E,i}, 0)\right\} \cdot \frac{(1 - r(S_{E,i}, X_{E,i})) \cdot r(S_{E,i}, X_{E,i})}{e(X_{E,i})}\right].
\end{aligned}$$

Using the same argument we can write (A.4) as

$$-\mathbb{E}_E\left[\left\{\mu_E(S_{E,i}, X_{E,i}, 0) - \mu_E(S_{E,i}, X_{E,i}, 1)\right\} \cdot \frac{(1 - r(S_{E,i}, X_{E,i})) \cdot r(S_{E,i}, X_{E,i})}{1 - e(X_{E,i})}\right].$$

Combining the results for (A.3) and (A.4) leads to

$$\mathbb{E}_E\left[\left\{\mu_E(S_{E,i}, X_{E,i}, 1) - \mu_E(S_{E,i}, X_{E,i}, 0)\right\} \cdot \frac{(1 - r(S_{E,i}, X_{E,i})) \cdot r(S_{E,i}, X_{E,i})}{e(X_{E,i}) \cdot (1 - e(X_{E,i}))}\right].$$

Collecting the last two terms, (A.5) and (A.6), we have

$$\begin{aligned}
&\mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i}, X_{E,i})|W_{E,i} = 1, X_{E,i}\right]\right] - \mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i})|W_{E,i} = 1, X_{E,i}\right]\right] \\
&+\mathbb{E}_E\left[\mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i})|W_{E,i} = 0, X_{E,i}\right]\right] - \mathbb{E}_E\left[\mathbb{E}_E\left[h_E(S_{E,i}, X_{E,i})|W_{E,i} = 0, X_{E,i}\right]\right] \\
&= \mathbb{E}_E\left[h_E(S_{E,i}, X_{E,i}) \cdot \frac{r(S_{E,i}, X_{E,i})}{e(X_{E,i})}\right] - \mathbb{E}_O\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{r(S_{E,i}, X_{E,i})}{e(X_{E,i})}\right] \\
&+\mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{1 - r(S_{E,i}, X_{E,i})}{1 - e(X_{E,i})}\right] - \mathbb{E}_E\left[h_O(S_{E,i}, X_{E,i}) \cdot \frac{1 - r(S_{E,i}, X_{E,i})}{1 - e(X_{E,i})}\right] \\
&= \mathbb{E}_E\left[\{h_E(S_{E,i}, X_{E,i}) - h_O(S_{E,i}, X_{E,i})\} \cdot \frac{r(S_{E,i}, X_{E,i}) - e(X_{E,i})}{e(X_{E,i}) \cdot (1 - e(X_{E,i}))}\right].
\end{aligned}$$

Combining the results for (A.3) and (A.4) with those for (A.5) and (A.6) then leads to the result in (ii). □

Proof for Theorem 3: The first representation of the efficiency bound without surrogacy is derived in Robins and Rotniztky (1995), Robins, Zhao and Rotnitzky (1995), and 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

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

Here $\hat{\mu}_E(s, w)$ is the average outcome for units with $S_{E,i} = s$ and $W_{E,i} = w$, and $\hat{\pi}_E(s|w) = P(S_{E,i} = s|W_{E,i} = w)$. Let $\hat{\pi}_E(s)$ be the fraction of units with $S_i = s$. Let $\pi_E(s|w)$ and $\pi_E(s)$ be the corresponding population probabilities, so that $\pi_E(s|1) = \pi_E(s) \cdot r(s)/p$.
We can write the difference between $\hat{\tau}$ and $\tau = \sum_{m=1}^{M} \pi_E(s_m|1) \cdot \mu_E(s_m, 1) - \sum_{m=1}^{M} \pi_E(s_m|0) \cdot \mu(s_m, 0)$ as

$$\hat{\tau} - \tau = \sum_{m=1}^{M} \hat{\pi}_E(s_m|1) \cdot (\hat{\mu}_E(s_m, 1) - \mu_E(s_m, 1)) - \sum_{m=1}^{M} \hat{\pi}_E(s_m|0) \cdot (\hat{\mu}_E(s_m, 0) - \mu_E(s_m, 0))$$

$$+ \sum_{m=1}^{M} (\hat{\pi}_E(s_m|1) - \pi_E(s_m|1)) \cdot \mu_E(s_m, 1) - \sum_{m=1}^{M} (\hat{\pi}_E(s_m|0) - \pi_E(s_m|0)) \cdot \mu_E(s_m, 0).$$

Up to the relevant order of approximation this is equal to

$$\hat{\tau} - \tau \approx \sum_{m=1}^{M} \pi_E(s_m|1) \cdot (\hat{\mu}_E(s_m, 1) - \mu_E(s_m, 1)) - \sum_{m=1}^{M} \pi_E(s_m|0) \cdot (\hat{\mu}_E(s_m, 0) - \mu_E(s_m, 0))$$

$$+ \sum_{m=1}^{M} (\hat{\pi}_E(s|1) - \pi_E(s|1)) \cdot \mu_E(s_m, 1) - \sum_{m=1}^{M} (\hat{\pi}_E(s|1) - \pi_E(s|0)) \cdot \mu_E(s_m, 0).$$

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

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

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

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

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

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

$$+\sum_{m=1}^{M}\frac{\pi_E(s_m)\cdot r(s_m)}{N\cdot p^2}\cdot(\mu_E(s_m,1)-\mu_1)^2+\sum_{m=1}^{M}\frac{\pi_E(s_m)\cdot r(s_m)}{N\cdot(1-p)^2}\cdot(\mu_E(s_m,0)-\mu_0)^2$$
$$=\frac{1}{N}\cdot\mathbb{E}_E\left[\sigma^2(S_{E,i})\cdot\left(\frac{r(S_{E,i})}{p^2}+\frac{1-r(S_{E,i})}{(1-p)^2}\right)\right.$$
$$\left.+\frac{r(S_{E,i})}{p^2}\cdot(\mu(S_{E,i})-\mu_1)^2+\frac{1-r(S_{E,i})}{(1-p)^2}\cdot(\mu(S_{E,i})-\mu_0)^2\right].$$

Now consider the case with surrogacy. The estimator now is

$$\hat{\tau}-\tau=\sum_{m=1}^{M}\hat{\pi}_E(s_m|1)\cdot(\hat{h}_E(s_m)-\mu_E(s_m,1))-\sum_{m=1}^{M}\hat{\pi}_E(s|0)\cdot(\hat{h}_E(s_m)-\mu_E(s_m,0))$$
$$+\sum_{m=1}^{M}(\hat{\pi}_E(s|1)-\pi_E(s|1))\cdot\mu_E(s_m,1))-\sum_{m=1}^{M}(\hat{\pi}_E(s|0)-\pi_E(s|0)\cdot\mu_E(s_m,0)),$$

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

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

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

$$\sum_{m=1}^{M}(\pi_E(s_m|1)-\pi_E(s_m|0))^2\cdot\mathbb{V}(\hat{h}_E(s_m))$$
$$=\sum_{m=1}^{M}\pi_E(s_m)^2\left(\frac{r(s_m)}{p}-\frac{1-r(s_m)}{1-p}\right)^2\cdot\frac{\sigma^2(s_m)}{N\cdot\pi_E(s_m)}$$
$$=\frac{1}{N}\sum_{m=1}^{M}\pi_E(s_m)\left(\frac{r(s_m)-p}{p\cdot(1-p)}\right)^2\cdot\sigma^2(s_m)$$
$$=\frac{1}{N}\sum_{m=1}^{M}\pi_E(s_m)\left(\frac{r(s_m)}{p}+\frac{1-r(s_m)}{1-p}-\frac{r(s_m)\cdot(1-r(s_m))}{p^2\cdot(1-p)^2}\right)\cdot\sigma^2(s_m)$$
$$=\frac{1}{N}\cdot\mathbb{E}_E\left[\sigma^2(S_{E,i})\cdot\left(\frac{r(S_{E,i})}{p^2}+\frac{1-r(S_{E,i})}{(1-p)^2}-\frac{r(S_{E,i})\cdot(1-r(S_{E,i}))}{p^2\cdot(1-p)^2}\right)\right.$$

Combining this with the last term leads to

$$\mathbb{V}_E(\hat{\tau})\approx\frac{1}{N}\cdot\mathbb{E}_E\left[\sigma^2(S_{E,i})\cdot\left(\frac{r(S_{E,i})}{p^2}+\frac{1-r(S_{E,i})}{(1-p)^2}-\frac{r(S_{E,i})\cdot(1-r(S_{E,i}))}{p^2\cdot(1-p)^2}\right)\right.$$

$$+\frac{r(S_{E,i})}{p^2}\cdot(h_E(S_{E,i})-\mu_1)^2+\frac{1-r(S_{E,i})}{(1-p)^2}\cdot(h_E(S_{E,i})-\mu_0)^2\Big].$$

□

The proof for Theorem 4 is similar and is omitted.

References

Abadie, A., and G. Imbens, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235-267.

Abadie, A., and G. Imbens, (2016), "Matching on the Estimated Propensity Score," *Econometrica*, Vol 84(2), 781-807.

Adams, K., A. Schatzkin, T. Harris, V. Kipnis, T. Mouw, R. Ballard-Barbash, A. Hollenbeck, and M. Leitzmann, (2006), "Overweight, Obesity, and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old", *New England Journal of Medicine*, Vol 355(8): 763-778.

s'Agostinho, R., M. Campbell, and G. Greenhouse, (2006), "Surrogate Markers: Back to the Future," (editorial) *Statistics in Medicine*, Vol. 25: 181-182.

Alonso, A., G. Molenberghs, H. Geys, M. Buyse, and T. Vangeneugden, (2006), "A Unifying Approach for Surrogate Marker Validation Based on Prentice's Criteria," *Statistics in Medicine*, Vol. 25: 205-221.

Athey, S., and S. Stern, (2002), "The impact of information technology on emergency health care reforms", *Rand Journal of Economics,*, Vol. 33: 399-432.

Belloni, A., V. Chernozhukov, and C. Hansen, (2014), "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81: 608-650.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner, (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer.

Begg, C., and D. Leung, (2000), "On the Use of Surrogate End Points in Randomized Trials," *Journal of the Royal Statistical Society, Series A*, 163(1): 15-28.

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

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*," *Quarterly Journal of Economics* 126, no. 4.

Ding, P., and J. Lu, (2015), "Principal Stratification Ananlysis Using Principal Scores," http://arxiv.org/pdf/1602.01196.pdf.

Fleming, T., and D. DeMets, (1996), "Surrogate End Points in Clinical Trials: Are We Being Misled," *Annals of Internal Medicine*, Vol. 125(7): 605-613.

Frangakis, C., and D. Rubin, (2002), "Principal Stratification," *Biometrics*, Vol (1): 21-29.

Freedman, D., (2008), "On Regression Adjustmens to Experimental Data," *Advances in Applied Mathematics* , Vol 30(6), 180-193.

Gelman, A., G. King, and . Liu, (1998), "Not Asked and Not Asnwered: Multiple Imputation for Multiple Surveys", *Journal of the American Statistical Association*, Vol. 93(443), 846-857.

Gilbert, P. and M. Hudgens, (2008), "Evaluating Candidate Principal Surrogate Endpoints," *Biometrics*, Vol. 64(4): 1146-1154.

Graham, B., C. Campos de Xavier Pinto, and D. Egel, (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economics and Statistics*, , Vol. (79), 10531079.

Graham, B., C. Campos de Xavier Pinto, and D. Egel, (2016), "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business and Economic Statistics*.

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

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189.

Holland, P., (1986), "Statistics and Causal Inference"(with discussion), *Journal of the American Statistical Association*, 81, 945-970.

Imbens, G., and D. Rubin, (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kang, J., and Schafer, J, (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22 (4), 523-539.

Van Der Laan, M., and M. Petersen, (2004), "Estimation of Direct and Indirect Causal Effects in Longitudinal Studies", U.C. Berkeley Division of Biostatistics Working Paper Series Paper 155.

Little, R., and D. Rubin, (1988), *Statistical Analysis with Missing Data*, Wiley.

NcGee, D., and the Diverse Populations Collaboration, (2004), "Body Mass Index and Mortality: A Meta-analysis Based on Person-level Data from Twenty-six Observational Studies", *Annals of Epidemiology*, Vol 15: 87-97.

Mealli, F., and A. Mattei, (2012), "A Refreshing Account of Principal Stratification," *International Account of Biostatistics*, 81(1): 1-17.

Morgan, S. and C. Winship, (2007), *Counterfactuals and Causal Inference*, Cambridge University Press, Cambridge.

Pearl, J., (2000, 2009), *Causality: Models, Reasoning and Inference*, Cambridge, Cambridge University Press.

Prentice, R., (1989), "Surrogate Endpoints in Clinical Trials: definition and Operational Criteria," *Statistics in Medicine*, Vol. 8: 431-440.

Rässler, S., (2002), *Statistical Matching*, Springer.

Rässler, S., (2004), "Data Fusion: Identification Problems, Validity, and Multiple Imputation," *Austrian Journal of Statistics*, 33, 153-171.

Ridder, G., and R. Moffitt, (2007), "The Econometrics of Data Combination," *Handbook of Econometrics*, Heckman and Leamer, eds., Vol 6B, 5469-5548.

Robins, J.M., Rotnitzky, A., Zhao, L-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106-121.

Rosenbaum, P.,, (1984), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment", *Journal of the Royal Statistical Society, Series A*, 147(5): 656-666.

Rosenbaum, P., (1995, 2002), *Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rubin, D., (1976), "Inference and Missing Data", *Biometrika*, Vol. 63(3): 581:592.

Rubin, D., (1986), "Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputation", *Journal of Business and Economic Statistics*, Vol. 4(1): 71-94.

Rubin, D. (2006), *Matched Sampling for Causal Effects*, Cambridge University Press, Cambridge.

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

VanderWeele, (2015), *Explanation in Causal Inference: Methods for Mediation and Interaction*, Oxford University Press.

Wager, S. and S. Athey, (2015), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," http://arxiv.org/pdf/1510.04342v2.pdf.

Weir, C., and R. Walley, (2006), "Statistical Evaluation of Biomarkers as Surrogate Endpoints: A Literature Review," *Statistics in Medicine*, Vol. 25: 183-203.

Xu, J., and S. Zeger, (2001), "The Evaluation of Multiple Surrogate Endpoints," *Biometrics*, 57(1): 81-87.