Learning plug-in surrogate endpoints for randomized experiments

Alessandro-Umberto Marguerittе[*1,2] **Ahmet Zahid Balcıoğlu**[*2]

Jesse H. Krijthe[3] **Dave Zachariah**[4] **Fredrik D. Johansson**[2]

[1]AstraZeneca
[2]Chalmers University of Technology and the University of Gothenburg
[3]TU Delft
[4]Uppsala University

Abstract

Surrogate endpoints are used in place of long-term outcomes in randomized experiments when observing the real outcome for a large enough cohort is prohibitively expensive or impractical. A short-term surrogate is good if the result of an experiment using the surrogate is predictive of the result of a hypothetical study using the real outcome. Much attention has been paid to formalizing this property in causal terms, but most criteria are unidentifiable and cannot be turned into practical algorithms for learning surrogate endpoints from data. To address this, we study plug-in composite surrogates, functions of post-treatment variables that may be substituted directly for the primary outcome in a randomized experiment. We propose two methods for learning plug-in surrogates that maximize effect predictiveness, and characterize the possibility of finding endpoints that yield unbiased effect estimates in representative scenarios. Finally, in both synthetic experiments with known effects and in data from a real-world experiment, we find that our method, based on directly modeling the surrogate effect, returns plug-in endpoints more predictive of the primary effect than established methods.

1 INTRODUCTION

In many experiments, such as a clinical trial or A/B tests, the long-term effects of interventions are difficult to estimate since the outcomes of interest rarely occur within the study period. For instance, nutrition science is notorious for its many examples (Yetley et al., 2017), where the risk of cardiovascular disease is an outcome of interest, but few study participants can maintain a prescribed diet long enough to see disease develop. Instead, experimentalists have turned to measurable *surrogates* of the primary outcomes—variables hoped to be informative of the efficacy of the intervention *and* yield reliable effect estimates on the primary outcome with tractable trial lengths and cohort sizes.

Surrogate endpoints have been identified, variously, by known mechanistic relationships (Fleming, 1996; Temple, 1999) or statistical associations (Prentice, 1989; Baker, 2018) with the primary outcome, as mediators of the causal effect (Chen et al., 2007; Robins and Greenland, 1992), or as variables on which the causal effect of the intervention is predictive of the causal effect on the primary outcome (Prentice, 1989; Chen et al., 2007; Meyvisch, 2020; Burzykowski et al., 2005). Formalizations of these conditions (Chen et al., 2007; Joffe and Greene, 2009; VanderWeele, 2013) have opened the possibility for the data-driven search for new surrogates based on observational (Athey et al., 2025) or past experimental data (Tripuraneni et al., 2024; Wang et al., 2022; Zhang et al., 2023), to design experiments.

Learning surrogate endpoints from observational data requires the causal identification (Pearl, 2009) of surrogacy criteria as learning objectives, as well as ensuring the transportability (Pearl and Bareinboim, 2011) of estimates to future experiments. However, most criteria are based on matching the *average* causal effects of surrogate and outcome (VanderWeele, 2013), quantities that are sensitive to distribution shift, and are insufficient for personalized decision-making. Under appropriate assumptions, this can be overcome by learning a set of nuisance functions that capture effect heterogeneity, confounding due to pre-treatment variables, and transportability, and using these for statistical estimates in the trial (Athey et al., 2025). However, using such functions as trial endpoints is undesirable, as they depend directly on pre-treatment variables.

In this work, we study learning *plug-in surrogates* from observational data—functions of only post-treatment variables that can be used in experiments or clinical trials as direct substitutes for the primary outcome. We contribute

- Desiderata and a corresponding learning objective for

[*]Equal contribution.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4272–4298.

plug-in surrogates, aimed at capturing as much heterogeneity in the causal effect on the outcome as possible.

- An analysis of scenarios where good plug-in surrogates exist, and lead to effect estimates that are unbiased for the primary outcome, and where they don't.
- Two algorithms designed to minimize our proposed objective and an empirical evaluation on synthetic and real-world data, highlighting their strengths and weaknesses compared to established methods.

We find that our algorithm, which directly models the surrogate effect by sampling counterfactual surrogate values from estimated densities, consistently returns plug-in endpoints that are more predictive of the primary-outcome effect than those of baseline methods.

2 PROBLEM SETUP & RELATED WORK

We study surrogates for the primary outcome $Y \in \mathbb{R}$ in a randomized experiment, aimed at estimating the causal effect of a binary intervention (treatment) T on Y in the context of a set of pre-treatment variables $X \in \mathbb{R}^k$. We focus on experiments where Y cannot be observed directly due to practical limitations, such as the outcome being too long-term to observe (e.g., cardiovascular disease or death), and we must rely on a set of post-treatment variables $S \in \mathbb{R}^d$ (e.g., metabolic markers) as candidate surrogates for Y.

Our goal is to *learn* a composite endpoint $f(S)$, a function $f : \mathbb{R}^d \to \mathbb{R}$ of the candidate surrogates S, that can serve *as a plug-in replacement for* Y in a typical randomized experiment. First, f is learned from retrospective data from an observational population, denoted $P = o$. Second, an experiment $(P = e)$ is performed, in which each instance i is randomly assigned a treatment $t_i \in \{0, 1\}$. The post-treatment surrogate vector $s_i \in \mathbb{R}^d$ is observed and used to compute $f_i = f(s_i)$, which is in turn used to estimate the treatment effect on f, as if f_i were the primary outcome y_i. The trial proceeds exactly as it would have if Y were observed, and conclusions about the effects of T on Y are drawn based on the effects of T on $f(S)$. For this to be a sound and effective strategy, $f(S)$ must satisfy the following desiderata:

D1 The intervention T has a causal effect on the composite endpoint $f(S)$ that is measurable in an experiment

D2 The causal effect of T on $f(S)$ is predictive of the causal effect of T on the primary endpoint Y, ideally capturing important effect heterogeneity

D3 The composite $f(S)$ is an interpretable function of post-treatment variables that is learned *before* the trial

The desiderata D1–D3 capture the requirements of many applied researchers who *want* to 1) find surrogates that are post-treatment variables with 2) similar effect as the main outcome, by 3) learning from historical data. However, existing approaches either fail to address domain requirements or are causally unsound. We show what it takes to solve this problem, when it is possible, and make the necessary assumptions explicit.

We expand on these desiderata, in relation to previous works. First, we let $S(0), S(1), Y(0), Y(1)$ denote the full vector of *potential outcomes* (Rubin, 1974) of S and Y, corresponding to interventions $t \in \{0, 1\}$ on T. The causal effects of T on the surrogate set and primary outcome are defined by

$$\Delta_S = S(1) - S(0) \quad \text{and} \quad \Delta_Y = Y(1) - Y(0)\,.$$

We refer to these as the *surrogate set effect* and *primary effect*, respectively. For a composite endpoint function $f : \mathbb{R}^d \to \mathbb{R}$, we define $\Delta_{f(S)} = f(S(1)) - f(S(0))$ and refer to this as the *surrogate effect*.

Regression of outcomes on post-treatment variables may return confounded plug-in surrogates.
A common strategy for finding plug-in surrogates among multiple candidate variables is to fit a regression of Y onto S, producing an estimate $f(S) \approx \mathbb{E}[Y \mid S]$ and using this as a composite score or to select a single variable as surrogate (Coley et al., 2020; Esposito et al., 2004; Martínez-González et al., 2015; Keys et al., 1984; Chan and Watts, 2006). Often, a linear model $f(S) = \theta^\top S$ is used either as the score itself, or to select a variable $f(S) = S_l$ where $l = \arg\max_j |\theta_j|$, based on the sizes of the fitted coefficients.

The strengths and weaknesses of the regression strategy are well known (Fleming and DeMets, 1996; Ciani et al., 2023; Baker and Kramer, 2003; Baker, 2018). Without confounding and with perfect mediation, i.e., when the causal graph contains only $T \to S \to Y$, it yields a surrogate whose effect $\tau_{f(S)}$ is predictive of the effect on the primary outcome. However, this precludes both direct effects of T on Y, and common causes of S and Y. If S and Y are both causally affected by variable X, as in Figure 1, the outcome regression strategy will lead to confounded effect estimates in general. We show how this bias emerges and give an example scenario in Appendix F.3. Moreover, when S contains variables that are predictive of Y but unaffected by T, variable selection or constrained/regularized models can fail, as they may favor associations that are irrelevant for the causal effect. To avoid this, we aim to learn composite surrogates f that *capture the causal effect of* T *on* Y (D1).

Learning to match average causal effects leads to underspecified surrogate endpoints.
The causal effect on a composite surrogate $f(S)$ is predictive (D2) of the primary effect if Δ_Y is informative of

$\Delta_{f(S)}$. A surrogate with this property enables consistent decision-making to improve Y, based on estimates of $\Delta_{f(S)}$ or its aggregates. If D2 is false, a trial may find no effect on the surrogate, even though there is an effect on the primary outcome. The reverse error can be equally costly, for example, if the effect on a surrogate does not replicate in a costly long-term trial where the main outcome is measured.

Regrettably, the joint distribution $p(\Delta_Y, \Delta_{f(S)})$ is unidentifiable (see Appendix B.1), as is e.g., the mutual information $I(\Delta_{f(S)}; \Delta_Y)$ (Meyvisch, 2020) and correlation $\rho(\Delta_{f(S)}, \Delta_Y)$, prohibiting their use in learning f. Even with infinitely many randomized trials, we could not learn or validate a surrogate $f(S)$ based on such criteria. In fact, most surrogacy criteria in the literature are unidentifiable (VanderWeele, 2013). Fortunately, most experiments have a more modest goal: to estimate the average (primary) treatment effect (ATE) in the experiment population,

$$\tau^e_Y = \mathbb{E}[\Delta_Y \mid P = e] \,.$$

The trial average surrogate effect $\tau^e_{f(S)} = \mathbb{E}[\Delta_{f(S)} \mid P = e]$ and observational-study counterparts τ^o_Y and $\tau^o_{f(S)}$ are defined analogously, the latter two conditioned on $P = o$. Henceforth, we let p^o, p^e and $\mathbb{E}^o, \mathbb{E}^e$ denote densities and expectations under the observational and experimental populations, respectively, leaving out conditioning on P.

Most surrogacy criteria are based on the possibility to recover the primary-outcome ATE using the surrogate (Prentice, 1989; Freedman et al., 1992; Frangakis and Rubin, 2002). For example, Chen et al. (2007) defined a *consistent* surrogate S (in the binary case) as a mediator for which the sign of the product $\tau^e_S \cdot \tau^e_{S \to Y}$ is equal to the sign of τ^e_Y, where $\tau^e_{S \to Y}$ is the average causal effect of S on Y. This rules out the "surrogate paradox", where T and S are positively associated, S and Y are positively associated, but T has a negative causal effect on Y (Chen et al., 2007). For recent works on evaluating surrogates, see Appendix F.2.

ATE-based criteria suggest searching for a composite $f(S)$ to match the trial ATEs of the surrogate and primary outcome: $\tau^e_{f(S)} \approx \tau^e_Y$. Finding such a function would allow its use as a plug-in replacement for Y and support *uniform* policy decisions based on experiment outcomes. However, this strategy has important caveats. First, it makes for a weak learning signal, leaving f woefully **underspecified**. For example, under mild assumptions, there is a linear surrogate function $f(S) = \alpha S_j$, of any single variable S_j affected by T, such that $\tau^e_{f(S)} = \tau^e_Y$ (see Appendix F.4). Second, it **does not support personalized decision-making**, as it does not capture effect heterogeneity. Finally, the ATE is **not transportable** in general, $\tau^e_Y \neq \tau^o_Y$, as the distribution of contexts X (e.g., population of subjects) varies, $p^e(X) \neq p^o(X)$, making the trial ATE difficult to learn in observational data.

The surrogate index addresses transportability and personalization but depends on pre-treatment variables.

To support personalized or contextual decision-making, it is suitable to estimate the *conditional average treatment effect* (CATE) with respect to strata $X = x$,

$$\tau^e_Y(x) = \mathbb{E}^e[\Delta_Y \mid X = x] \,,$$

and act according to the preferred treatment for each stratum. We define $\tau^e_{f(S)}(x)$, $\tau^o_Y(x)$, and $\tau^o_{f(S)}(x)$ analogously. To simplify notation, X will take both the roles of potential confounding variables and treatment effect moderators. In general, these could be separated into two distinct variables.

The *surrogate index*, proposed by Athey et al. (2025), removes several of the limitations of ATE-matching strategies by estimating and transporting the nuisance function

$$h^o(x, s) \coloneqq \mathbb{E}^o[Y \mid X = x, S = s] \,. \tag{1}$$

Under ignorability, Prentice's surrogacy condition, and comparability, all to be defined formally later, it holds that

$$\tau^e_Y(x) = \mathbb{E}^e[h^o(x, S) \mid x, T = 1] - \mathbb{E}^e[h^o(x, S) \mid x, T = 0]$$

and, by definition, $\tau^e_Y = \mathbb{E}^e[\tau^e_Y(X)]$. Thus, both the trial ATE and CATE can be identified by learning $h^o(x, s)$ in observational data and averaging its predictions in the trial.

However, several properties of the surrogate index $h^o(x, s)$ make it **unsuitable as a plug-in surrogate** in clinical trials (D3). First, $h^o(x, s)$ depends directly on pre-treatment variables X that cannot mediate the effect of T on Y. This is generally not accepted for clinical trial endpoints (Friedman et al., 2015), as pre-treatment variables are also used for trial inclusion criteria, and could be used to manipulate the estimated effect. Second, using the surrogate index relies on statistical quantities only estimable in trial data, such as the trial propensity $p(P = e \mid X, S)$, which prevents using $h^o(x, s)$ in place of Y in a difference-in-means estimate of τ^e_Y. Third, a surrogate must be such that a governing authority will accept the evidence provided by a trial using it as an endpoint. Most likely, f needs to belong to an interpretable model family, but the surrogate index analysis does not consider approximate estimates $h^o(x, s) \approx \mathbb{E}^o[Y \mid x, s]$.

3 LEARNING COMPOSITE ENDPOINTS

To challenge the limitations of existing strategies, we propose learning composite plug-in surrogate endpoints by searching for a function $f(S)$ with *conditional* trial treatment effect most similar to the primary outcome, by solving

$$\underset{f \in \mathcal{F}}{\text{minimize}} \;\; R^e_\tau(f) \coloneqq \mathbb{E}^e_X \left[\left(\tau^e_Y(X) - \tau^e_{f(S)}(X) \right)^2 \right] , \tag{2}$$

where $\tau^e_{f(S)}$ is the conditional surrogate effect given by

$$\tau_{f(S)}(x) \coloneqq \mathbb{E}\left[f(S(1)) - f(S(0)) \mid X = x \right] .$$

The function f will be learned *before* the experiment starts from two sources of retrospective, typically observational, data sampled from distributions $p^o(X, T, S)$ and $p^o(X, S, Y)$, respectively. We do *not* assume that all four variables are observed together. For instance, the first data could be the second-phase trials for a new treatment, while the latter could come from health records of past patients.

An endpoint $f(S)$ from an interpretable[1] model family $\mathcal{F}$ with small trial CATE error $R^e_\tau(f)$ satisfies our desiderata by optimizing for the predictability of the effect of T on Y (D1–D2), and serving as a plug-in surrogate based only on post-treatment variables, learned completely before the experiment (D3). Additionally, it supports **personalization** by being trained to capture treatment effect heterogeneity, and **transportability** by optimizing for use in $p^e(X)$ directly.

Next, we prove the identifiability of $R^e_\tau(f)$, give strategies to minimize it, and discuss limitations of plug-in surrogates, paying attention to the conditions under which the method yields unbiased estimates of the primary treatment effect.

3.1 IDENTIFYING THE TRIAL CATE ERROR

To solve (2) before the experiment starts, we must rely on statistical parameters estimable from observational data. Both $\tau_Y(x)$ and $\tau_{f(S)}(x)$ are unknown interventional quantities which must be causally *identified* to avoid confounding and other biases. For this, we make classical identifying assumptions on the support and causal (in)dependences of variables, analogous to Athey et al. (2025). We also give a summary of comparisons of our settings and the different goals of our respective works in Section F.1.

Assumption 1 (Ignorability)**.** *The context X gives conditional exchangeability for S and Y in $P = o$, $\forall t \in \{0, 1\}$,*

$$Y(t) \perp\!\!\!\perp T \mid X, P = o \quad S(t) \perp\!\!\!\perp T \mid X, P = o \,,$$

and satisfies positivity (common support),

$$\forall x : p^o(x) > 0 \implies p^o(T = t \mid X = x) > 0 \,.$$

Endpoints satisfy consistency, $Y = Y(T), S = S(T)$.

Exchangeability can be guaranteed by graphical criteria, such as the backdoor criterion (Pearl, 2009), applied to the causal graph of the problem's variables. The four causal graphs in Figure 1 all guarantee exchangeability. We return to these in Section 3.3 to discuss the biases of plug-in surrogates. For more discussion on graphical criteria for surrogacy, see Lauritzen (2003); VanderWeele (2013).

S is a set of variables, each of which may be associated with a different set of confounders. Thus, justifying conditional exchangeability of $S(t)$ is, in general, more difficult than for a single outcome, Y. Moreover, performing an experiment intervening on T can only remove confounding from the association between T and Y, but not between S and Y, and it may not be possible to intervene on S itself.

[1]We use the term "interpretable" in the way established by e.g., Rudin (2019), typically transparent models with short descriptions. The right choice of model class is domain-dependent.

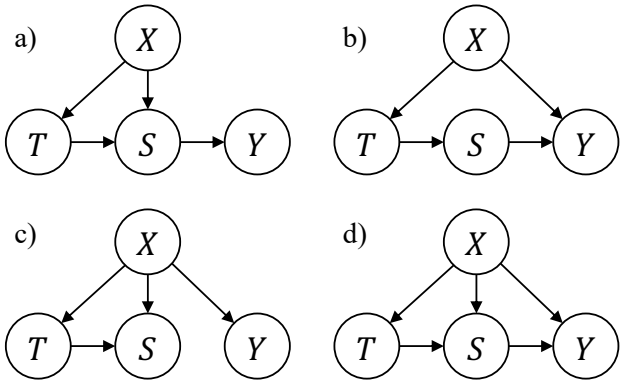


Figure 1: Special cases of causal graphs compatible with Assump. 1–3, for which the CATE error (2) is identifiable.

Under Assumption 1, standard arguments yield that

$$\begin{aligned}\mu^o_Y(x,t) &:= \mathbb{E}^o[Y(t) \mid X = x] = \mathbb{E}^o[Y \mid X = x, T = t] \\ &= \mathbb{E}^o_S[\mathbb{E}^o_Y[Y \mid S, X = x, T = t] \mid X = x, T = t]\end{aligned}$$

and $p^o(S(t) \mid X = x) = p^o(S \mid X = x, T = t)$. For most definitions of surrogacy (Athey et al., 2025; Chen et al., 2007; Prentice, 1989), it is additionally assumed that

Assumption 2 (Prentice surrogacy)**.** *The surrogate set S and context X render the outcome independent of treatment,*

$$Y \perp\!\!\!\perp T \mid X, S, P = p \quad \textit{for} \quad p \in \{o, e\}.$$

Assumption 2 is not trivial to satisfy and is not possible to test without observing the variables X, T, S, Y together. Instead, it requires careful problem-specific justification, and is subject to debate among practitioners (Prentice, 1989). For it to hold, important backdoor paths should be blocked by X—loosely, X should account for important confounders. Additionally, S should account for the causal effect of T on Y. This rules out T having a direct effect on Y, or an effect mediated by an unobserved variable (see Figure 2). This is typically argued by reasoning about known mechanistic relationships, and whether S contains proxies for these. This is more plausible when S is a well-chosen set of multiple relevant variables, rather than a single factor.

Under Assumptions 1–2, the conditional expected potential outcomes $\mu^o_Y(x, t)$ can be identified from $p^o(X, T, S)$ and $p^o(X, S, Y)$ since the conditioning on T in the inner expectation above can be dropped. It follows trivially that $\tau^o_{f(S)}(x)$ and $\tau^o_Y(x)$ are identified as well. To identify Objective (2), we additionally need to relate p^e and p^o. Broadly, this is a problem of transportability (Pearl and Bareinboim,

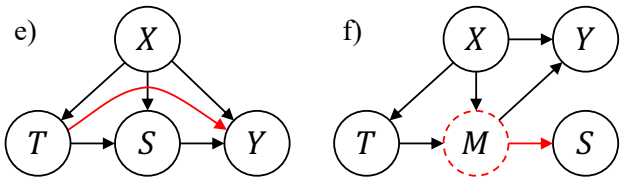


Figure 2: Challenging cases for surrogacy. In case e), there is a direct $T \to Y$ effect which prevents finding an optimal plug-in surrogate. In case f), the true mediator M is unobserved, and S are only proxies for M.

2011), typically solved by imposing assumptions on shared conditionals between p^e and p^o, so also here.

Assumption 3 (Transportability)**.** *For the experiment and observational populations. p^e, p^o, it holds for all x, t, s that*

$$p^e(S \mid X = x, T = t) = p^o(S \mid X = x, T = t)$$
$$p^e(Y \mid X = x, S = s) = p^o(Y \mid X = x, S = s)\,.$$

The second equality is assumed by Athey et al. (2025) as "comparability". Here, the assumption on the distribution of S is made to support learning $f(S)$ ahead of the experiment.

Proposition 1. *Under Assump. 1–3, with $h^o(x, s)$ in (1),*

$$\begin{aligned}\tau^e_Y(x) = \tau^o_Y(x) &= \mathbb{E}^o_S[h^o(x, S) \mid X = x, T = 1] \\ &\quad - \mathbb{E}^o_S[h^o(x, S) \mid X = x, T = 0] \\ \tau^e_{f(S)}(x) = \tau^o_{f(S)}(x) &= \mathbb{E}^o_S[f(S) \mid X = x, T = 1] \\ &\quad - \mathbb{E}^o_S[f(S) \mid X = x, T = 0]\,.\end{aligned}$$

We give a proof in Appendix B following from standard arguments. To fully identify (2), the trial population $p^e(X)$ or the density ratio $p^e(X)/p^o(X)$ must be known, to account for differences between the observational and experiment cohorts. The surrogate index strategy of Athey et al. (2025) does not rely on pre-experiment access to this information since only the surrogate index $h^o(X, S)$ is transferred from the observational study. This is not an option under desideratum D3. Moving forward, we assume that $p^e(X)/p^o(X)$ is known a priori to an accuracy sufficient to find an approximate minimizer of (2). This is possible, for example, when the main distributional difference between experimental and observational cohorts is due to inclusion criteria I applied for the experiment, as is standard practice in clinical trials. With $\mathbb{1}[x \in I]$ indicating that x satisfies the inclusion criteria, it is plausible that $p^e(x) \propto p^o(x)\mathbb{1}[x \in I]$.

Finally, building on Proposition 1, under Assumptions 1–3, we can identify the trial CATE error as follows

$$R^e_\tau(f) = \mathbb{E}^o_X\Big[\frac{p^e(X)}{p^o(X)}\big(\mathbb{E}^o_S[h^o(X, S) - f(S) \mid X, T = 1] - \mathbb{E}^o_S[h^o(X, S) - f(S) \mid X, T = 0]\big)^2\Big]. \quad (3)$$

which is estimable from $p^o(X, T, S)$ and $p^o(X, S, Y)$. Moving forward, we drop the superscript from $h^o(x, s)$.

3.2 MINIMIZING THE EXPECTED CATE ERROR

Minimizing (3) requires estimating $R^e_\tau(f)$ in a way that is amenable to optimization of f. Below, we present two strategies for this. Both start by fitting an estimate $\hat{h}(x, s) \approx \mathbb{E}^o[Y \mid X = x, S = s]$ by regression on an observational data set $D^o_y = \{(x'_j, s'_j, y'_j)\}_{j=1}^{m_y}$ drawn from $p^o(X, S, Y)$.

Estimation through surrogate sampling

Using $\hat{h}(x, s)$ in place of $h(x, s)$ in (3), the expression depends only on variables (x, s, t) and can be estimated using data set $D^o_t = \{(x_j, t_j, s_j)\}_{j=1}^{m_t}$. However, each sample holds only an observation of the factual potential outcome $S_i(t_i)$ for a given x, but not of the counterfactual $S_i(1 - t_i)$, making it difficult to estimate the inner expectations.

If an accurate model $\hat{p}(S \mid x, t) \approx p(S(t) \mid x)$ is available, the inner expectations of (3) can, in principle, be computed exactly, but only for trivially small problems or simple distributions (such as when f, h are linear, see Section 3.3). Instead, we propose sampling potential outcomes $S(0)$ and $S(1)$ from $\hat{p}(S \mid x, t)$ to yield augmented data points $(x_i, s_i, t_i, \hat{s}(0)^1_i, ..., \hat{s}(0)^L_i, \hat{s}(1)^1_i, ..., \hat{s}(1)^L_i,)$, with L samples $\hat{s}(t)^l_i$ of each potential outcome, used to estimate

$$\mathbb{E}_S[\hat{h}(x_i, S) \mid X = x_i, T = t] \approx \frac{1}{L}\sum_{l=1}^{L} \hat{h}(x_i, \hat{s}(t)^l_i),$$

and $\mathbb{E}_S[f(S) \mid x_i, t]$, analogously. Plugging these estimates into (3) for $t \in \{0, 1\}$, and averaging over $p^e(x)$, we have an estimator $\hat{R}^e_\tau(f)$ of the full trial CATE risk. We refer to the algorithm that fits $f(S)$ by minimizing this estimate as **Surrogate sampling**, described in Algorithm 1. In the stated version, we ignore population differences $p^e(x) \neq p^o(x)$. Such differences can be accommodated by re-weighting the risk average with the density ratio $p^e(x)/p^o(x)$.

The minimizers f^* of (3) are translation invariant, that is, $R^e_\tau(f) = R^e_\tau(f + b)$ for any constant b, and may not align in scale with the primary outcome, Y. This can be easily post-calibrated using samples of Y or the fitted function $\hat{h}$.

Estimating conditional densities $p(S \mid x, t)$ for continuous S can be difficult and relies on the availability of large data sets (Papamakarios et al., 2021). Discretizing the variable set, the conditional density may be represented by a probability table, but results in expensive computation and statistical challenges. For example, when S is composed of only six covariates, approximating each covariate by ten bins results in estimating $p(S = s|X = x, T = t)$ for 10^6 different configurations s, for each value of x.

Algorithm 1 Learning plug-in composite endpoints

Input: Observational treatment data $D_t^o = \{(x_j, t_j, s_j)\}_{j=1}^{m_t}$ and outcome data $D_y^o = \{(s'_i, x'_i, y'_i)\}_{i=1}^{m_y}$, model class $\mathcal{F}$
Fit outcome model $\hat{h}(x, s) \approx \mathbb{E}[Y \mid x, s]$ to D_y^o
if Surrogate sampling **then**
  Fit surrogate models $\hat{g}(x, t) \approx p(S \mid x, t)$ for $t \in \{0, 1\}$
  for $j = 1, ..., m, l = 1, ..., L$ **do**
    Draw $\hat{s}(t)_j^l \sim \hat{g}(x_j, t)$ for $t \in \{0, 1\}$
  end for
  $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{j=1}^{m_t} \Big(\sum_{t=0}^{1} (-1)^t \sum_{l=1}^{L} [\hat{h}(x_j, \hat{s}(t)_j^l) - f(\hat{s}(t)_j^l)] \Big)^2$
else if Bound regression **then**
  Fit propensity score $\hat{e}(x) \approx p(T = 1 \mid x)$ to D_t^o
  Fit surrogate score $\hat{\rho}(x, s) \approx p(T = 1 \mid x, s)$ to D_t^o
  $\hat{w}^2(x, s) = \big(\hat{\rho}(x, s)/\hat{e}(x) - (1 - \hat{\rho}(x, s))/(1 - \hat{e}(x))\big)^2$
  $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{j=1}^{m_t} \hat{w}^2(x_j, s_j)(\hat{h}(x_j, s_j) - f(s_j))^2$
end if
return $\hat{f}$
Note: If $p^e(x) \neq p^o(x)$, the objectives should be weighted by $p^e(x)/p^o(x)$.

Minimizing an upper bound on the CATE risk

To avoid sampling-based estimation or costly numerical integration of Objective 2, we can minimize an upper bound of the objective, resulting in a weighted regression problem.

Proposition 2. *For any surrogate function f, under Assumptions 1–3, the trial CATE risk is bounded as*

$$R_\tau^o(f) \leq \mathbb{E}_{X,S}[w^2(X, S)(h(X, S) - f(S))^2] \leq 2\mathbb{E}_{X,S}[w^2(X, S)(Y - f(S))^2] + 2\sigma^2, \quad (4)$$

where $w^2(x, s) = \left(\frac{p(T=1|s,x)}{p(T=1|x)} - \frac{p(T=0|s,x)}{p(T=0|x)} \right)^2$. The second inequality holds if the outcome has exogenous, centered, additive noise: $Y = h(X, S) + \epsilon$ where $\mathbb{E}[\epsilon] = 0$, $Var(\epsilon) \leq \sigma^2$ and $\epsilon \perp\!\!\!\perp S, X, T$. If $p^o(X)/p^e(X)$ is known, using weights $\tilde{w}^2(x, s) = \frac{p^o(x)}{p^e(x)} w^2(x, s)$ yields a bound on $R_\tau^e(f)$.

A proof is given in Appendix B.4. There, we also derive analogous results to (4) with non-squared, additive weights $w^+(x, s) = \frac{p(T=1|s,x)}{p(T=1|x)} + \frac{p(T=0|s,x)}{p(T=0|x)}$ for the squared loss and absolute difference weights $w^1(x, s) = \left| \frac{p(T=1|s,x)}{p(T=1|x)} - \frac{p(T=0|s,x)}{p(T=0|x)} \right|$ for the L_1 loss. A downside of the w^+ weights is that they are uniform and uninformative when $p(T = 1 \mid x) \equiv 0.5$, e.g., in a randomized experiment.

Remarks. Minimizing (4) amounts to solving a weighted regression problem with x, s as input and $h(x, s)$, or y, as label.[2] The weight $w^2(x, s)$ places emphasis on points (x, s) where S is informative of T conditioned on X, which is precisely points (x, s) where s is more affected by the treatment assignment. This stands in contrast to standard outcome-on-surrogate regression (with weights $w(x, s) \equiv 1$), which has no such preference. An algorithm based on minimizing the bound in (4) using an estimate $\hat{h}$ in place of h is presented in Algorithm 1 as "**Bound regression**".

3.3 THE BIASES OF PLUG-IN SURROGATES

It is instructive to ask: When is there *any* plug-in surrogate $f(S)$ such that $\tau_{f(S)}(x) = \tau_Y(x)$?[3] And if there is one, can it be learned using one of our methods? If $\|\tau_S\| > 0$, as described previously, it is always possible to find f such that the ATEs are matched, $\tau_{f(S)} = \tau_Y$, but do our methods return such an f? To begin to answer this, we examine the four cases in Figure 1 where $R_\tau^e(f)$ is identifiable. For proofs and discussions on each case, see Appendix C.

Case a) Confounded surrogate mediator

Here, $Y \perp\!\!\!\perp X \mid S$, and so $h(x, s) = \mathbb{E}[Y \mid X = x, S = s] = \mathbb{E}[Y \mid S = s]$. Consequently, $\tau_Y(x) = \mathbb{E}_S[\mathbb{E}[Y \mid S]|X = x, T = 1] - \mathbb{E}_S[\mathbb{E}[Y \mid S]|X = x, T = 0] = \tau_{f(S)}(x)$ for a perfect regression $f(s) = \mathbb{E}[Y \mid S = s]$. This f both minimizes (2) and yields an unbiased estimate of $\tau_Y(x)$ and τ_Y. Moreover, *any* weighted regression, with full weight support over x, s and sufficient functional capacity, will learn this function, including the weights $w^2(x, s)$.

Case b) Confounded outcome, surrogate mediator

Here, $S \perp\!\!\!\perp X \mid T$ and S carries no information about effect heterogeneity due to X. Out of the four cases, this is the worst for *any* plug-in surrogate in terms of matching CATE, whether found by minimization of (2) or not. In general, for all f, $\exists x, s : f(s) \neq h(x, s)$ and, unless $p^o(s \mid T = 1, x) = p^o(s \mid T = 0, x)$, implying that S is unaffected by T, $\exists x : \tau_{f(S)}(x) \neq \tau_Y(x)$. Thus, plug-in surrogate estimates of the primary-outcome CATE are generally biased. It is however, possible, to match the ATE, and bound regression using the weights w^+ or w^1 does yield $\tau_{f(S)} = \tau_Y$.

Case c) Confounded surrogate, no effect

Here, the primary-outcome effects $\tau_Y(x) = \tau_Y = 0$. Any function f such that $f(S) \perp\!\!\!\perp T \mid X$ will yield $\tau_{f(S)}(x) = 0$, such as a constant $f(S) = c$. Thus, an unbiased plug-in surrogate exists and would be found by surrogate sampling, as it minimizes (2), but is not guaranteed to be found by bound regression or outcome regression since $Y \not\perp\!\!\!\perp T \mid S$ due to confounding by X.

[2] However, using Y as label requires access to samples of all variables (x, t, s, y) at once, which may not be available.

[3] We can drop the population superscripts here as $\tau^e(x) = \tau^o(x)$ under Assumptions 1–3.

Case d) Confounded surrogate and outcome

Here, the extent to which τ_Y can be matched by a function $f(S)$ is determined by several factors, such as the functional form of $\mathbb{E}[Y \mid s, x]$ and the extent to which information in x that alters $\tau_Y(x)$ is captured by s. In general, no plug-in surrogate is unbiased. We give the regression weights unbiased for the ATE in Appendix C.4, but these are potentially negative and are not suitable for weighted regression. An important special case is when the contributions of X and S are additive in Y. We highlight this in the case when the contribution of S is linear, and show the advantage of our learning objective over outcome regression even when X and S are non-additive in Y.

The linear case. When the functions f and h are both linear with respect to s, that is, $h(x, s) = \phi(x) + \beta_h^\top s + b_h$ and $f(s) = \beta_f^\top s + b_f$, the trial CATE risk simplifies to

$$R^e_\tau(f) = \mathbb{E}^e_X[((\beta_h - \beta_f)^\top \tau_S(x))^2],$$

as shown in Appendix B.3. Taking $\beta_f = \beta_h$ yields $R^e_\tau(f) = 0$, and hence, the plug-in surrogate $f(S)$ that minimizes (2) is unbiased for both $\tau_Y(x)$ and τ_Y, as is the surrogate index estimator (Athey et al., 2025). However, standard (unweighted) outcome regression with a linear model $\theta^\top s$ is biased in general, since $\mathbb{E}[Y|S = s] = \mathbb{E}[\phi(X) \mid s] + \beta_h^\top s + b_h$ and $\mathbb{E}_{S(1)}[\mathbb{E}_X[\phi(X) \mid S(1)]] \neq \mathbb{E}_{S(0)}[\mathbb{E}_X[\phi(X) \mid S(0)]]$ without further assumptions.

Dominant baseline variation. Even if X and S are not additive in Y, our learning objective (2) can be more advantageous than using outcome regression alone. For instance, when h is composed of baseline and effect terms, $h(x, s) = \mathbb{E}[Y \mid x, s] = \phi(x) + \Delta(x, s)$, and $\phi(x)$ dominates its variability, outcome regression estimating $\mathbb{E}[Y \mid s]$ would be dominated by the term $\mathbb{E}[\phi(x) \mid s]$ and variable selection methods would select surrogates that explain the noise that do not depend on the treatment T. In contrast, our learning objective learns $\mathbb{E}[\Delta(x, s) \mid s]$, more accurately capturing the treatment effect.

Choosing moderators X to personalize policies. Based on the four cases above, the best-case scenario to learn a plug-in surrogate $f(S)$ is when X and S are tightly linked, to the point that $Y \perp\!\!\!\perp X \mid S$. We can design successful surrogate studies by *choosing* both X and S to achieve this. For example, by letting X include pre-treatment measurements of the variables in S, and S include follow-up measurements of the confounders in X makes it likely that the information in X relevant to predict the effect on the outcome Y is contained in S. We remind the reader that X is composed of two types of variables: (i) adjustment variables to remove confounding, (ii) effect moderators used to personalize decisions. We treat them as the same set here, but they need not be. If a confounder is unhelpful for personalization, it need not be included in the CATE conditioning set, and it removes one more source of potential effect variance for S to explain.

4 EXPERIMENTS

We evaluate our proposed methods[4] in comparison with standard baselines and recent methods for surrogate learning in a suite of synthetic experiments and on data from the IHDP randomized controlled trial ("IHDP", 1990; Martin et al., 2008). For details on the experimental setup, see Appendix D.

Methods. As baselines, we include outcome regressions, estimating $\mathbb{E}[Y \mid S = s]$, as well as a single-surrogate regression (Reg.-sel.-reg.), learned by fitting a regression, selecting the variable with the largest absolute coefficient, and regressing only on that. We also include the surrogate index $h(x, s) = \hat{\mathbb{E}}[Y|x, s]$ as a comparison that depends directly on pre-treatment variables, x. All baselines come in linear and tree model variants, with gradient boosting regression for the surrogate index. For nonlinear models, we do a 5-fold cross validation and grid search to select hyperparameters.

We implement **Bound regression** using weighted linear regression. For fitting propensity weights $e(x)$ and surrogate score $\rho(x, s)$, we use logistic regression, fitted using a L_2 regularization $\alpha = 1$. For **Surrogate sampling**, we first fit nuisance random forest models for $\hat{\mathbb{E}}[S|x, T = t]$ and sample $L = 50$ residuals of the fitted model via bootstrapping to simulate the conditional $\hat{p}(S|x, T = t)$. Then, we fit a linear regression model on the potential outcomes $\hat{s}(0), \hat{s}(1)$ sampled from the conditional (see Algorithm 1), using PyTorch (Paszke et al., 2019). For IHDP, we additionally implement a tree variant of **Bound regression** using the `PyDL8.5` library (Aglin et al., 2021), which learns globally optimal trees, using our proposed weighted squared-error objective. This variant is included to explore interpretable composite surrogates in real-data applications, where transparent decision rules and explicit surrogate partitions facilitate substantive interpretation of the learned surrogates.

For each surrogate f, we estimate the trial surrogate ATE $\tau^e_{f(S)}$ using the average treatment-group outcomes in the trial $\hat{\mathbb{E}}^e[f(S) \mid T = 1] - \hat{\mathbb{E}}^e[f(S) \mid T = 0]$. We estimate the surrogate CATE $\tau^e_{f(S)}(x)$, using a T-learner, fitting linear models $g_t(x) \approx \mathbb{E}_S[f(S)|x, t]$ for $t \in \{0, 1\}$ and defining $\tau^e_{f(S)}(x) = g_1(x) - g_0(x)$. We compare this to an estimate using the simulated potential outcomes $s(1)$ and $s(0)$ and found strong agreement in results (see appendix Figure 4).

[4] Code and instructions for reproducing the experiments are available at https://github.com/Healthy-AI/causal-plugin-surrogates.

Table 1: Synthetic scenarios a–d), matching the graphs in Figure 1, see Appendix D. Results are averaged over linear and nonlinear scenarios within cases. We report 95% confidence intervals computed using the standard error across runs. In Cases b–c), the CATE is constant and R^2 is uninformative. †The surrogate index depends directly on pre-treatment variables.

Composite Scenarios	**Case a)** $\overline{ATE}$=3.74		**Case b)** $\overline{ATE}$=2.12	**Case c)** $\overline{ATE}$=0	**Case d)** $\overline{ATE}$=5.67		**Case d) Linear** $\overline{ATE}$=18.54	
Method	MAE $\hat{\tau}\downarrow$	$R^2\hat{\tau}(x)\uparrow$	MAE $\hat{\tau}\downarrow$	MAE $\hat{\tau}\downarrow$	MAE $\hat{\tau}\downarrow$	$R^2\hat{\tau}(x)\uparrow$	MAE $\hat{\tau}\downarrow$	$R^2\hat{\tau}(x)\uparrow$
Reg.-sel.-reg. (lin)	2.22± 0.16	0.27± 0.23	0.72±0.02	0.32± 0.01	2.35± 0.13	−24.0± 16.0	1.98± 0.06	0.85± 0.01
Reg.-sel.-reg. (tree)	1.23± 0.15	0.41± 0.25	0.71±0.07	0.26± 0.01	1.53± 0.13	−23.9± 2.53	1.40± 0.03	0.81± 0.01
Outcome reg. (lin)	1.53± 0.15	0.40± 0.02	0.14±0.06	0.55±0.02	1.22±0.12	0.51± 0.15	1.63± 0.05	0.90± 0.00
Outcome reg. (tree)	0.87± 0.14	0.33± 0.03	0.40±0.06	0.41±0.03	0.66±0.12	0.40± 0.18	0.93± 0.07	0.89± 0.01
Surrogate index† (lin)	1.68± 0.15	0.31± 0.03	0.16±0.06	0.64± 0.03	1.17±0.13	0.51± 0.13	0.04± 0.04	0.96± 0.00
Surrogate index† (tree)	0.86± 0.14	0.33± 0.5	0.40±0.06	0.35±0.03	0.67±0.12	0.01± 0.40	0.16± 0.08	0.90± 0.01
Surrogate index† (hgb)	0.34± 0.12	0.54± 0.32	0.11±0.06	0.25±0.03	0.21±0.11	0.61± 0.15	0.07± 0.05	0.95± 0.00
Bound reg. (lin)	1.69± 0.15	0.37± 0.26	0.17±0.06	0.35±0.02	1.33±0.13	0.48± 0.14	1.11± 0.05	0.93± 0.00
Bound reg. (tree)	0.68± 0.13	0.44± 0.29	0.31±0.06	0.32±0.02	0.81±0.11	−2.10± 1.85	1.27± 0.06	0.89± 0.01
Surrogate sampl. (lin)	0.50± 0.13	0.65± 0.27	0.09±0.05	0.20±0.01	0.40±0.12	0.66± 0.15	0.03± 0.03	0.97± 0.00

Evaluation metrics. We evaluate the quality of learned surrogates $f(S)$ by the mean absolute error (MAE $\hat{\tau}$) of surrogate-based estimates $\hat{\tau}^e_{f(S)}$ with respect to the primary-outcome ATE τ^e_Y, and the coefficient of determination R^2 of the surrogate CATE $\hat{\tau}^e_{f(S)}(x)$ with respect to $\tau^e_Y(x)$, averaged across seeds and scenarios. For details, see Appendix D.3. In synthetic data, we use ATE and CATE computed using simulated potential outcomes as ground-truth. In IHDP, we use difference-in-means estimates of the ATE as ground-truth. The uncertainty of each metric is assessed using 95% percentile confidence intervals computed using bootstrap resampling.

4.1 SYNTHETIC EXPERIMENTS

We generate synthetic observational and trial cohorts from structural causal models over (X, T, S, Y), obeying scenarios a-d) in Figure 1, where $T \in \{0, 1\}$, $X \in \mathbb{R}^k$ for $k \in \{2, 5\}$ and the surrogate set S is composed of three variables, $S_{\text{med}} \in \mathbb{R}^3$, $S_{\text{leaf}} \in \mathbb{R}^2$, and $S_{\text{proxy}} \in \mathbb{R}^2$ where $S_{\text{med}}, S_{\text{leaf}}$ are affected by T and $S_{\text{med}}, S_{\text{proxy}}$ are causes of Y. In the observational cohort, treatments are assigned based on X, but randomized uniformly in the experimental cohort; otherwise data generation follows the same process. Across scenarios, we include linear and nonlinear outcome mappings and variations in treatment overlap. Each scenario is repeated with multiple random seeds. For a full description, including structural equations, see Appendix D. The results of the synthetic experiments are presented in Table 1.

Surrogate sampling achieved the best overall results. Across scenarios and metrics, directly minimizing the estimated CATE trial error through surrogate sampling (Algorithm 1) with a linear model achieved the best results for both ATE and CATE among methods that return a plug-in surrogate function of only post-treatment variables, nearly matching the performance of the strongest baseline, a surrogate index using a more flexible histogram gradient boosting model which uses both X and S as inputs.

Bound regression comparable to outcome regression. In most cases, we see little benefit of the bound regression algorithm, with Case d) Linear as a notable exception. Here, the linear bound regression outperforms the (unweighted) linear outcome regression. This is expected as there exists an unbiased linear plug-in surrogate in this case, as demonstrated by the Surrogate sampling method.

Composite surrogates outperform single-variable surrogates. Across scenarios, seeds, and model classes, the Regress-select-regression baselines fared the worst, indicating that a single-variable surrogate is insufficient to predict the effect of the intervention on the primary outcome.

4.2 REAL-WORLD EXPERIMENT: IHDP

The Infant Health and Development Program (IHDP)[5] was a randomized longitudinal multi-site trial of early childhood intervention for low birth weight, premature infants ("IHDP", 1990). Baseline covariates X include information about the mother and child measured at birth, and the surrogate vector S consists of post-baseline measures of child development, collected up to 24 months after birth. The primary outcome Y is the Stanford–Binet IQ score at 36 months corrected age. This outcome is not possible to measure before 24 months, as children of this age cannot complete an IQ score (Laurent et al., 1992), necessitating the use of a surrogate to estimate the effect at this time.

[5]A semi-synthetic version of the IHDP study (Hill, 2011) has been used widely in the machine learning community, but this is *not* used here. We use the original data from the randomized trial.

Table 2: ATE recovery on IHDP. The ground-truth ATE, $\tau^e = 6.47$. ATE and MAE are reported as bootstrap means with 95% percentile CIs. †The surrogate index depends directly on pre-treatment variables.

Method	$\widehat{\text{ATE}}$ (95% CI)	MAE (95% CI)
Reg-Sel-Reg (lin)	6.11 (2.20, 10.34)	1.78 (0.13, 4.65)
Reg-Sel-Reg (tree)	7.55 (3.28, 11.34)	1.92 (0.09, 4.95)
Outcome reg. (lin)	5.41 (0.89, 9.37)	1.98 (0.09, 5.68)
Outcome reg. (tree)	7.82 (3.56, 11.80)	2.05 (0.14, 5.43)
Surrogate index† (lin)	4.79 (-0.06, 9.59)	2.32 (0.07, 6.53)
Surrogate index† (tree)	6.27 (2.15, 10.35)	1.70 (0.08, 4.91)
Bound reg. (lin)	4.69 (0.91, 8.48)	2.13 (0.12, 5.56)
Bound reg. (tree)	6.74 (2.94, 9.90)	1.52 (0.07, 3.69)
Bound reg. (DL8.5)	6.60 (2.84, 10.24)	1.56 (0.06, 4.14)
Surrogate sampl. (lin)	5.49 (2.75, 8.03)	1.35 (0.04, 3.72)

We restrict the cohort to complete cases on $\{T, Y, X, S\}$ and create a 70/30 split stratified by treatment status with a fixed random seed. The 70-split is used to learn surrogate models as the $P = o$ cohort; the held-out split is used exclusively for trial validation, $P = e$. Table 2 reports ATE recovery on the held-out split for the fixed-propensity specification. We show the experimental ATE, the model-implied ATE, and the absolute ATE error. Variable definitions and measurement details are provided in Appendix D.4.

Broad agreement on the benefits of the intervention. All methods estimate a positive overall treatment effect on IQ at 36 months, consistent with the experimental benchmark $\tau^e = 6.47$. Bootstrap means are generally close to the ground-truth ATE, though uncertainty remains substantial across specifications. Most methods identify the `Bayley_MDI` score as a prominent surrogate. Importantly, all methods would have led to the same positive conclusion about the treatment effect using data available at 24 months rather than waiting for the 36 month IQ outcome, highlighting the practical value of early surrogate information.

Minimizing the trial CATE error yields the most predictive surrogates. Surrogate sampling achieves the smallest bootstrap MAE, followed by tree bound regression and DL8.5. These methods combine relatively small bias with tighter uncertainty bands compared to simpler regression baselines and the surrogate index. Selection and outcome-regression approaches display wider dispersion and moderately larger errors, while the linear surrogate index exhibits the largest deviation from the experimental ATE.

Overall, as in the synthetic experiments, the surrogate sampling strategy performs particularly well in recovering the global ATE, suggesting that explicitly modeling the surrogate distribution can improve finite-sample stability when rich post-treatment information is available.

5 DISCUSSION

We have studied the learning of plug-in composite surrogate endpoints, functions of post-treatment variables that act as substitutes for unobserved long-term outcomes in randomized experiments. We proposed two algorithms based on minimizing estimates or bounds of the CATE error—the expected squared difference between the conditional average treatment effect estimated using the surrogate and using the primary outcome. In evaluations on multiple simulated scenarios and data from a real-world randomized experiment, we find that our method, based on directly modeling the surrogate effect by sampling surrogate counterfactuals, returns stronger plug-in endpoints than baselines.

Our long-term goal with this work is to control the quality of decisions made based on the results of a trial using the surrogate endpoint in place of the primary outcome. With nothing else impacting utility, this comes down to matching the signs of the average treatment effect for a uniform policy, or the signs of the conditional effect for a personalized one. For future work, learning surrogates that yield a point-wise bound on the conditional average treatment effects is an interesting direction that would support this strategy.

Our current implementation of the surrogate sampling method is limited to linear models. In future work, we will explore a tree-based variant of surrogate sampling, which requires a new tree algorithm as the objective depends on both potential outcomes. This variant will benefit from the low bias of our best-performing method, and the capability to capture nonlinear treatment effect heterogeneity.

A question left open by this work is learning from surrogate variables, which are not mediators, but proxies for an unobserved mediator (such as in case (f) in Figure 2). Using such variables as surrogates violates Prentice surrogacy (Assumption 2) and requires a new identification strategy. Finally, we hope to characterize the statistical properties of trial estimates based on learned plug-in surrogates. This would not just make certain randomized trials feasible, but require fewer participants.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. JHK was partially supported by the IDEA League Fellowship.

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and 2024-03903.

References

Gaël Aglin, Siegfried Nijssen, and Pierre Schaus. Pydl8. 5: a library for learning optimal decision trees. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5222–5224, 2021.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Review of Economic Studies*, page rdaf087, 2025.

Stuart G. Baker. Five criteria for using a surrogate endpoint to predict treatment effect based on data from multiple previous trials. *Statistics in medicine*, 2018. doi: 10.1002/sim.7561.

Stuart G Baker and Barnett S Kramer. A perfect correlate does not a surrogate make. *BMC medical research methodology*, 3(1):16, 2003.

Tomasz Burzykowski, Marc Buyse, and Geert Molenberghs. *The evaluation of surrogate endpoints*, volume 427. Springer, 2005.

DC Chan and GF Watts. Apolipoproteins as markers and managers of coronary risk. *Journal of the Association of Physicians*, 99(5):277–287, 2006.

Hua Chen, Zhi Geng, and Jinzhu Jia. Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5):919–932, 2007.

Oriana Ciani, Anthony M Manyara, Philippa Davies, Derek Stewart, Christopher J Weir, Amber E Young, Jane Blazeby, Nancy J Butcher, Sylwia Bujkiewicz, An-Wen Chan, et al. A framework for the definition and interpretation of the use of surrogate endpoints in interventional trials. *EClinicalMedicine*, 65, 2023.

Nicola Coley, Marieke P Hoevenaar-Blom, Jan-Willem van Dalen, Eric P Moll van Charante, Miia Kivipelto, Hilkka Soininen, Sandrine Andrieu, Edo Richard, and PRODEMOS consortium, the preDIVA study group, the MAPT/DSA group, and the HATICE consortium. Dementia risk scores as surrogate outcomes for lifestyle-based multidomain prevention trials—rationale, preliminary evidence and challenges. *Alzheimer's & Dementia*, 16(12): 1674–1685, 2020.

Michael R Elliott, Anna SC Conlon, Yun Li, Nico Kaciroti, and Jeremy MG Taylor. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics*, 16(2):400–412, 2015.

Katherine Esposito, Raffaele Marfella, Miryam Ciotola, Carmen Di Palo, Francesco Giugliano, Giovanni Giugliano, Massimo D'Armiento, Francesco D'Andrea, and Dario Giugliano. Effect of a mediterranean-style diet on endothelial dysfunction and markers of vascular inflammation in the metabolic syndrome: a randomized trial. *Jama*, 292(12):1440–1446, 2004.

Thomas R Fleming. Surrogate endpoints in clinical trials. *Drug Information Journal*, 30(2):545–551, 1996.

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

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

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

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*. Springer, 2015.

Peter B Gilbert and Michael Hudgens. Evaluating causal effect predictiveness of candidate surrogate endpoints. *Biometrics*, 2006.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

"IHDP". Enhancing the outcomes of low-birth-weight, premature infants: A multisite, randomized trial. *JAMA*, 263(22):3035–3042, 06 1990. ISSN 0098-7484. doi: 10.1001/jama.1990.03440220059030.

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

Ancel Keys, Alessandro Menotti, Christ Aravanis, Henry Blackburn, Bozidar S Djordevič, Ratko Buzina, AS Dontas, Flaminio Fidanza, Martti J Karvonen, Noboru Kimura, et al. The seven countries study: 2,289 deaths in 15 years. *Preventive medicine*, 13(2):141–154, 1984.

Jeff Laurent, Mark Swerdlik, and Mary Ryburn. Review of validity research on the stanford-binet intelligence scale. *Psychological assessment*, 4(1):102, 1992.

Steffen L Lauritzen. Graphical models for surrogates. *Bull. Int. Statist. Inst*, 60:144–147, 2003.

Anne Martin, Jeanne Brooks-Gunn, Pamela Klebanov, Stephen L Buka, and Marie C McCormick. Long-term maternal effects of early childhood intervention: Findings from the infant health and development program

(ihdp). *Journal of Applied Developmental Psychology*, 29(2):101–117, 2008.

Miguel A Martínez-González, Jordi Salas-Salvadó, Ramón Estruch, Dolores Corella, Montse Fitó, Emilio Ros, Predimed Investigators, et al. Benefits of the mediterranean diet: insights from the predimed study. *Progress in cardiovascular diseases*, 58(1):50–60, 2015.

Paul Meyvisch. *Surrogate marker evaluation in clinical trials using methods of causal inference*. PhD thesis, KU Leuven, 2020.

Geert Molenberghs, Ariel Alonso Abad, and Wim Van der Elst. The statistical evaluation of surrogate endpoints in clinical trials. In *Biostatistics in Biopharmaceutical Research and Development: Clinical Trial Analysis, Volume 2*, pages 243–286. Springer, 2024.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 247–254, 2011.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

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

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Robert Temple. Are surrogate markers adequate to assess cardiovascular disease drugs? *Jama*, 282(8):790–795, 1999.

Nilesh Tripuraneni, Lee Richardson, Alexander D'Amour, Jacopo Soriano, and Steve Yadlowsky. Choosing a proxy metric from past experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5803–5812, 2024.

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

Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4100–4109, 2022.

Elizabeth A Yetley, David L DeMets, and William R Harlan Jr. Surrogate disease markers as substitutes for chronic disease outcomes in studies of diet and chronic disease relations. *The American journal of clinical nutrition*, 106 (5):1175–1189, 2017.

Vickie Zhang, Michael Zhao, Anh Le, Nathan Kallus, et al. Evaluating the surrogate index as a decision-making tool using 200 a/b tests at netflix. *arXiv preprint arXiv:2311.11922*, 2023.

A NOTATION

Table 3: A summary of used notations throughout the paper.

Random variables	
X	Pre-treatment variables
T	Treatment variable $T \in \{0, 1\}$
$Y, Y(t)$	Observed outcome and potential (interventional) outcomes
$S, S(t)$	Surrogate variables and surrogate potential outcomes
Δ_Y, Δ_S	$Y(1) - Y(0)$, $S(1) - S(0)$
U	Unobserved confounders
S_{med}	Mediator surrogates, $(T \to S \to Y)$
S_{leaf}	Leaf surrogates (only $T \to S$)
S_{proxy}	Proxy surrogates (only $S \to Y$)
$\tau_Y(x)$	Conditional average treatment effect (CATE) for Y
$\tau_{f(S)}(x)$	CATE estimate of a "plug-in" surrogate function f
$\tau_S(x)$	CATE for the surrogate set S
$\mu_Y(x, t)$	Conditional outcomes $\mathbb{E}[Y(t)\|X = x]$ for Y
Observations and constants	
$w(x, s)$	Weights of any weighted regression
$w^2(x, s)$	Weights of the **Bound regression**, $\left(\frac{p(T=1\|s,x)}{p(T=1\|x)} - \frac{p(T=0\|s,x)}{p(T=0\|x)}\right)^2$
$w^+(x, s)$	Weights of the **Bound regression**, $\frac{p(T=1\|s,x)}{p(T=1\|x)} + \frac{p(T=0\|s,x)}{p(T=0\|x)}$
$w^1(x, s)$	Weights of the **Bound regression**, $\left\|\frac{p(T=1\|s,x)}{p(T=1\|x)} - \frac{p(T=0\|s,x)}{p(T=0\|x)}\right\|$
$\rho(s, x)$	Surrogate score $p(T = 1\|x, s)$
$e(x)$	Propensity score, $p(T = 1 \mid x)$
$\eta(x, s)$	Difference of densities $\frac{p(T=1\|s,x)}{p(T=1\|x)} - \frac{p(T=0\|s,x)}{p(T=0\|x)}$
$\pi(x, s)$	Difference of densities $p(s \mid x, T = 1) = p(s \mid x, T = 0)$
$R^o_\tau(f)$, $R^e_\tau(f)$	Expected risk of the CATE estimate for f in the observational and trial data
$p^o(x)$, $p^e(x)$,	Marginal distribution of the covariates in the observational data and trial respectively
τ^o, τ^e, μ^o, μ^e	Observational and trial counterparts of treatment effects and expected outcomes
$\mathbb{E}^o$, $\mathbb{E}^e$	Observational and trial counterparts of expectations
D^o_t, D^o_y	Observational treatment and outcome data respectively
m_t, m_y	Number of samples for treatment and outcome data respectively
x_i	Observed pretreatment variables
t_i	Observed treatments
s_i	Observed surrogates
y_i	Observed outcomes
k	Number of pretreatment variables
d	Number of surrogate variables
L	Number of drawn samples in the **Surrogate sampling** methods
Functions	
$f(s)$	"Plug-in" surrogate functions learned via **Surrogate sampling** and **Bound regression** methods
$h(x, s)$	Surrogate index $h := \mathbb{E}[Y\|X = x, S = s]$

B PROOFS

B.1 ON THE IDENTIFIABILITY OF SURROGATE EFFECTS

Definitions of surrogacy that depend on knowing the joint distribution of surrogate and primary-outcome effects, $\Delta_S = S(1) - S(0)$ and $\Delta_Y = Y(1) - Y(0)$, will not in general be identifiable from observational or experimental data. It is well-known that even the joint distribution $p(Y(0), Y(1))$ is not identifiable other than under strong assumptions, and nor is $p(Y(1) - Y(0))$. Thus, neither is $p(Y(1) - Y(0), S(1) - S(0))$.

If there exists a variable X that captures non-treatment-related influence on Y such that it renders $Y(1), Y(0)$ conditionally independent, i.e., $Y(0) \perp\!\!\!\perp Y(1) \mid X$, identification of $p(Y(0), Y(1))$ is possible, and $p(Y(0), Y(1)) = \sum_{x \in X} p(x) p\left(Y(1) \mid x\right) p\left(Y(0) \mid x\right)$. However, for many causal models, such an X cannot exist, such as in structural causal models with additive (unobservable) noise, since this noise inextricably links $Y(0)$ and $Y(1)$.

It is, however, always possible to identify $p(Y(t) \mid X)$ and $p(S(t) \mid X)$ for some moderator X, provided that a sufficient adjustment set exists, which can be used to define an approximation of the joint distribution of effects. First, let

$$\tilde{p}(y_1, y_0 \mid x) = p(Y(1) = y_1 \mid x) p(Y(0) = y_0 \mid x)$$
$$\tilde{p}(s_1, s_0 \mid x) = p(S(1) = s_1 \mid x) p(S(0) = s_0 \mid x)$$

and then define a pseudo-joint distribution

$$\tilde{p}_X(\Delta_Y = \delta_y, \Delta_S = \delta_s) \coloneqq \sum_{\substack{x, y_0, y_1, s_0, s_1: \\ y_1 - y_0 = \delta_y \\ s_1 - s_0 = \delta_s}} p(x) \tilde{p}(y_1, y_0 \mid x) \tilde{p}(s_1, s_0 \mid x).$$

If X (however unlikely) renders all potential outcomes conditionally independent, this quantity coincides with the true joint distribution of surrogate and main effects. However, we do not pursue identification of $\tilde{p}_X(\Delta_Y, \Delta_S)$.

B.2 PROOF OF PROPOSITION 1

Proposition 1 (Restated). Under Assumptions 1–3,

$$\begin{aligned} \tau_Y^e(x) = \tau_Y^o(x) &= \mathbb{E}_S^o[h^o(x, S) \mid X = x, T = 1] \\ &\quad - \mathbb{E}_S^o[h^o(x, S) \mid X = x, T = 0] \\ \tau_{f(S)}^e(x) = \tau_{f(S)}^o(x) &= \mathbb{E}_S^o[f(S) \mid X = x, T = 1] \\ &\quad - \mathbb{E}_S^o[f(S) \mid X = x, T = 0]\,. \end{aligned}$$

Proof. We begin with $\tau_Y^o(x)$ given by

$$\begin{aligned} \tau_Y^o(x) &= \mathbb{E}[\Delta_Y | X = x, P = o] = \mathbb{E}^o[Y(1) - Y(0) \mid X = x] && (5) \\ &\underset{(a)}{=} \mathbb{E}^o[Y \mid X = x, T = 1] - \mathbb{E}^o[Y \mid X = x, T = 0] && (6) \\ &= \mathbb{E}_S^o[\mathbb{E}^o[Y \mid S, X = x, T = 1] - \mathbb{E}_S^o[\mathbb{E}^o[Y \mid S, X = x, T = 0]] && (7) \\ &\underset{(b)}{=} \mathbb{E}_S^o[\mathbb{E}^o[Y \mid S, X = x] \mid X = x, T = 1] - \mathbb{E}_S^o[\mathbb{E}^o[Y \mid S, X = x] \mid X = x, T = 0] && (8) \\ &\underset{(c)}{=} \mathbb{E}_S^o[h^o(x, S) \mid X = x, T = 1] - \mathbb{E}_S^o[h^o(x, S) \mid X = x, T = 0], && (9) \end{aligned}$$

where we use Assumption 1 in (a) and Assumption 2 in (b), and (c) follows from the definition of h^o. Combining (8) with Assumption 3 yields $\tau_Y^e(x) = \tau_Y^o(x)$. Similarly for $\tau_{f(S)}^o(x)$ we have

$$\begin{aligned} \tau_{f(S)}^o(x) &= \mathbb{E}_S^o[f(S) \mid X = x, T = 1] - \mathbb{E}_S^o[f(S) \mid X = x, T = 0] \\ &= \mathbb{E}_S^e[f(S) \mid X = x, T = 1] - \mathbb{E}_S^e[f(S) \mid X = x, T = 0] = \tau_{f(S)}^e(x). \end{aligned}$$

□

B.3 THE TRIAL CATE RISK IN THE LINEAR CASE

When the functions f and h are linear with respect to s, i.e. $h(x,s) = \phi(x) + \beta_h^\top s, f(s) = \beta^\top s$, the trial CATE risk simplifies to a linear function of the surrogate CATE:

$$\begin{aligned}
R_\tau^e(f) &= \mathbb{E}_X^e[(\mathbb{E}_S[h(X,S) - f(S) \mid X, T=1] \\
&\qquad - \mathbb{E}_S[h(X,S) - f(S) \mid X, T=0])^2] \\
&= \mathbb{E}_X^e[(\mathbb{E}_S[\phi(X) + \beta_h^\top S - \beta_f^\top S \mid X, T=1] \\
&\qquad - \mathbb{E}_S[\phi(X) + \beta_h^\top S - \beta_f^\top S \mid X, T=0])^2] \\
&= \mathbb{E}_X^e[((\beta_h - \beta_f)^\top \mathbb{E}_S[S \mid X, T=1] \\
&\qquad - (\beta_h - \beta_f)^\top \mathbb{E}_S[S \mid X, T=0])^2] \\
&= \mathbb{E}_X^e[((\beta_h - \beta_f)^\top \tau_S(x))^2].
\end{aligned}$$

□

B.4 PROOF OF PROPOSITION 2

Proposition 2 (Restated). For any surrogate function f, it holds under Assumptions 1–3 that

$$\begin{aligned}
R_\tau^o(f) &\leq \mathbb{E}_{X,S}[w^2(X,S)(h(X,S) - f(S))^2] \\
&\leq 2\mathbb{E}_{X,S}[w^2(X,S)(Y - f(S))^2] + 2\sigma^2
\end{aligned} \tag{10}$$

where $w^2(x,s) = \left(\frac{p(T=1|s,x)}{p(T=1|x)} - \frac{p(T=0|s,x)}{p(T=0|x)}\right)^2$, and the second inequality holds if the outcome has exogenous, centered, additive noise: $Y = h(X,S) + \epsilon$ where $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) \leq \sigma^2$ and $\epsilon \perp\!\!\!\perp S, X, T$. If $p^o(X)/p^e(X)$ is known, the same expression with the weight $\tilde{w}^2(x,s) = \frac{p^o(x)}{p^e(x)} w^2(x,s)$ yields a bound on $R_\tau^e(f)$.

Proof. We leave out superscripts indicating the population, as the proof holds for either cohort as long as it is used consistently on both the LHS and RHS.

Under Assumptions 1–2, we have consistency $S = S(T)$ and exchangeability on S, $S(t) \perp\!\!\!\perp T \mid X$, as well as $Y = Y(t)$, and $Y(t) \perp\!\!\!\perp T \mid X$, and $Y \perp\!\!\!\perp T \mid X, S$. Then, we can prove

$$\begin{aligned}
\mathbb{E}[Y(t) \mid X = x] &= \mathbb{E}[Y(t) \mid X = x, T = t] && (11)\\
&= \mathbb{E}[Y \mid X = x, T = t] && (12)\\
&= \mathbb{E}_{S(t)}[\mathbb{E}[Y \mid X = x, S(t), T = t] \mid X = x, T = t] && (13)\\
&= \mathbb{E}_S[\mathbb{E}[Y \mid X = x, S, T = t] \mid X = x, T = t] && (14)\\
&= \mathbb{E}_S[\mathbb{E}[Y \mid X = x, S] \mid X = x, T = t]\,. && (15)
\end{aligned}$$

Where we have used in (14) that

$$p(S(t) = s \mid X = x) = p(S = s \mid X = x, T = t)\,.$$

under the given assumptions. Let $h(x,s) \coloneqq \mathbb{E}[Y \mid X = x, S = s]$. Then,

$$\begin{aligned}
(\tau_Y(x) - \tau_{f(S)}(x))^2 &= (\mathbb{E}[h(x,S(1)) - f(S(1)) \mid X = x] - \mathbb{E}[h(x,S(0)) - f(S(0)) \mid X = x])^2 \\
&= (\mathbb{E}_S[(h(x,S) - f(S)) \mid X = x, T = 1] - \mathbb{E}_S[(h(x,S) - f(S)) \mid X = x, T = 0])^2 \\
&= \left(\mathbb{E}[\frac{p(S \mid T=1, x)}{p(S \mid x)}(h(x,S) - f(S)) \mid X = x] - \mathbb{E}[\frac{p(S \mid T=0, x)}{p(S \mid x)}(h(x,S) - f(S)) \mid X = x]\right)^2 \\
&= \left(\mathbb{E}\left[\left(\frac{p(T=1 \mid S, x)}{p(T=1 \mid x)} - \frac{p(T=0 \mid S, x)}{p(T=0 \mid x)}\right)(h(x,S) - f(S)) \mid X = x\right]\right)^2 \\
&\leq \mathbb{E}[w^2(x,s)(h(x,S) - f(S))^2 \mid X = x]
\end{aligned}$$

where $w^2(x,s) = \left(\frac{p(T=1|S,x)}{p(T=1|x)} - \frac{p(T=0|S,x)}{p(T=0|x)}\right)^2$, and so

$$\mathbb{E}_X[(\tau_Y(x) - \tau_{f(S)}(X))^2] \leq \mathbb{E}_{X,S}[w^2(X,s)(h(X,S) - f(S))^2] \,.$$

To arrive at the result using Y instead of $h(X,S)$, we exploit the assumption of additive, centered, exogenous noise ϵ, and that $Y = h(X,S) + \epsilon$ in this case. For any weighting function,

$$\begin{aligned}
\mathbb{E}_{X,S}[w(X,S)(h(X,S) - f(S))^2] &= \mathbb{E}_{X,S,Y}[w(X,S)(Y - f(S) + h(X,S) - Y)^2] \\
&\leq 2\mathbb{E}_{X,S,Y}[w(X,S)[(Y - f(S))^2 + (h(X,S) - Y)^2]] \\
&\leq 2\mathbb{E}_{X,S,Y}[w(X,S)[(Y - f(S))^2]] + 2\sigma^2
\end{aligned}$$

by the assumption of bounded noise from the statement. □

We can prove a similar result by bounding the square earlier,

$$\begin{aligned}
\frac{1}{2}(\tau_Y(x) - \tau_{f(S)}(x))^2 &= \frac{1}{2}(\mathbb{E}[h(x,S(1)) - f(S(1)) \mid X = x] - \mathbb{E}[h(x,S(0)) - f(S(0)) \mid X = x])^2 \\
&\leq \mathbb{E}[h(x,S(1)) - f(S(1)) \mid X = x]^2 + \mathbb{E}[h(x,S(0)) - f(S(0)) \mid X = x]^2 \\
&\leq \mathbb{E}_{S(1)}[(h(x,S(1)) - f(S(1)))^2 \mid X = x] + \mathbb{E}_{S(1)}[(h(x,S(0)) - f(S(0)))^2 \mid X = x] \\
&\leq \mathbb{E}_S[(h(x,S) - f(S))^2 \mid X = x, T = 1] + \mathbb{E}_S[(h(x,S) - f(S))^2 \mid X = x, T = 0] \\
&= \mathbb{E}\left[\frac{p(S \mid T = 1, x)}{p(S \mid x)}(h(x,S) - f(S))^2 \mid X = x\right] \\
&\quad + \mathbb{E}\left[\frac{p(S \mid T = 0, x)}{p(S \mid x)}(h(x,S) - f(S))^2 \mid X = x\right] \\
&= \mathbb{E}\left[\left(\frac{p(T = 1 \mid S, x)}{p(T = 1 \mid x)} + \frac{p(T = 0 \mid S, x)}{p(T = 0 \mid x)}\right)(h(x,S) - f(S))^2 \mid X = x\right] \\
&= \mathbb{E}\left[w^+(x,s)(h(x,S) - f(S))^2 \mid X = x\right]
\end{aligned}$$

where $w^+(x,s) = \frac{p(T=1|S,x)}{p(T=1|x)} + \frac{p(T=0|S,x)}{p(T=0|x)}$. It follows immediately that

$$\mathbb{E}_X[(\tau_Y(x) - \tau_{f(S)}(x))^2] \leq 2\mathbb{E}_{X,S}[w^+(x,s)(h(x,S) - f(S))^2] \,.$$

However, the weights w^+ are all uniform if $e(x) = p(T = 1 \mid x) = 0.5$, that is, if there is no confounding.

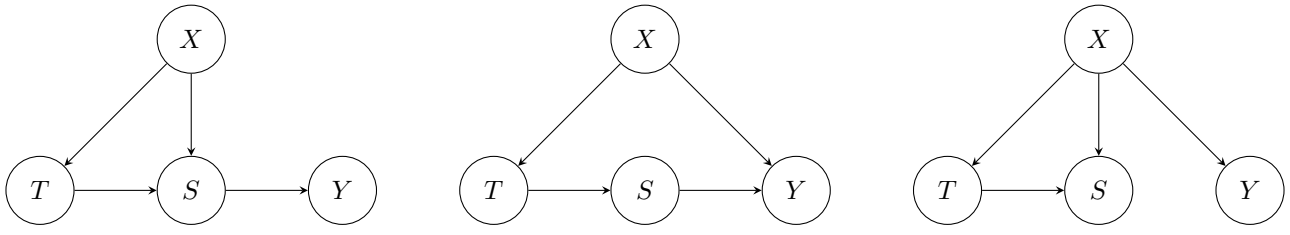
Using a very similar strategy, we can also prove a result for the L1-loss.

$$\begin{aligned}
|\tau_Y(x) - \tau_{f(S)}(x)| &= \left|\int_s \left(\frac{p(T = 1 \mid S, x)}{p(T = 1 \mid x)} - \frac{p(T = 0 \mid S, x)}{p(T = 0 \mid x)}\right) p(s \mid x)(h(x,s) - f(s))ds\right| \\
&= |\mathbb{E}_S[\eta(x,S)(h(x,S) - f(S)) \mid X = x]| \\
&\leq \mathbb{E}_S[w^1(x,S)|h(x,S) - f(S)| \mid X = x]
\end{aligned}$$

with $\eta(x,s) = \frac{p(T=1|S,x)}{p(T=1|x)} - \frac{p(T=0|S,x)}{p(T=0|x)}$, and $w^1(x,s) = |\eta(x,s)|$, which means

$$\mathbb{E}_X[|\tau_Y(X) - \tau_{f(S)}(X)|] \leq \mathbb{E}_{X,S}[w^1(x,s)|h(x,S) - f(S)|] \,.$$

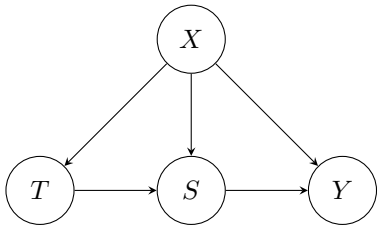
C EXAMPLES AND COUNTEREXAMPLES



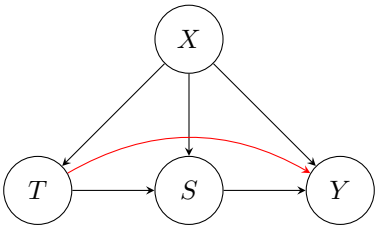
(a) Confounded surrogate mediator

(b) Confounded outcome

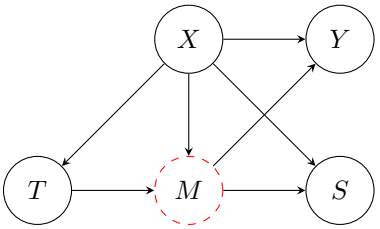
(c) Confounded surrogate, no effect



(d) Confounded surrogate and outcome



(e) Direct, unmediated effect



(f) Unobserved mediator

Figure 3: Six scenarios for surrogate learning.

Figure 3 illustrates the causal graphs of six surrogate learning scenarios. We go through the properties of cases a-e) below. In all of the cases illustrated, for any function $f(S)$

$$\tau_{f(S)}(x) = \mathbb{E}[f(S) \mid X = x, T = 1] - \mathbb{E}[f(S) \mid X = x, T = 0]$$

We also note that a convenient form to express $R^e_\tau(f)$ is

$$R^e_\tau(f) = \mathbb{E}^o_X \left[\frac{p^e(X)}{p^o(X)} \Big(\int_s \pi(x, s)(h(X, S) - f(S))ds \Big)^2 \right],$$

where

$$\pi(x, s) \coloneqq p(s \mid x, T = 1) - p(s \mid x, T = 0) \,.$$

C.1 A) CONFOUNDED SURROGATE MEDIATOR

In this case,

$$\mu_Y(x, t) = \mathbb{E}[Y(t) \mid X = x] = \mathbb{E}_{S(t)}[\mathbb{E}[Y \mid S(t)] \mid X = x] = \mathbb{E}_S[\mathbb{E}[Y \mid S] \mid X = x, T = t]$$

and with $g(s) = \mathbb{E}[Y \mid S = s]$, the CATE on Y is equal to

$$\tau_Y(x) = \mathbb{E}_S[g(S) \mid T = 1, x] - \mathbb{E}_S[g(S) \mid T = 0, x]$$

which is non-zero in general. Moreover, there exists an $f(S)$ such that $\tau_Y(x) = \tau_{f(S)}(x)$ for all x, and that $f = g + c$ for any constant c. Thus, $f = g + c$ are also the minimizers of (2). $f = g$ is also the minimizer of any weighted regression objective $\mathbb{E}[w(X, S)(h(X, S) - f(S))^2]$ when $h(X, S) = g(S)$.

C.2 B) CONFOUNDED OUTCOME

In this case,

$$\mu_Y(x, t) = \mathbb{E}[Y(t) \mid X = x] = \mathbb{E}_{S(t)}[\mathbb{E}[Y \mid S(t), X = x, T = t] \mid X = x] = \mathbb{E}_S[\mathbb{E}[Y \mid S, X = x] \mid X = x, T = t]$$

and with $h(x, s) = \mathbb{E}[Y \mid S = s, X = x]$,

$$\tau_Y(x) = \mathbb{E}_S[h(x, S) \mid T = 1, x] - \mathbb{E}_S[h(x, S) \mid T = 0, x]$$

which is non-zero in general. In this case, there *may not exist* a function $f(S)$ such that $\tau_Y(x) = \tau_{f(S)}(x)$ for all x, since X carries information about Y not mediated through S. In fact, $\mathbb{E}[f(S) \mid X = x, T = t] = \mathbb{E}[f(S) \mid T = t]$ is constant with respect to X since $S \perp\!\!\!\perp X \mid T$. Under mild assumptions, a minimizer of (2) yields the ATE,

$$\forall f \in f^* : \tau_{f(S)}(x) = \mathbb{E}_X[\mathbb{E}[h(X,S) \mid T = 1, X] - \mathbb{E}[h(X,S) \mid T = 0, X]] = \tau_Y \ .$$

We can find the minimizer f^* by setting the derivative of any weighted regression to zero, conditioned on s,

$$\mathbb{E}_X[2w(X,s)(h(X,s) - f(s)) \mid S = s] = 0 \Longrightarrow f^*(s) = \frac{\mathbb{E}_X[w(X,s)h(X,s) \mid S = s]}{\mathbb{E}_X[w(X,s) \mid S = s]}.$$

This expression is not constant with respect to s in general. Then

$$\begin{aligned}
\tau_{f^*(S)}(x) &= \mathbb{E}[f^*(S) \mid T = 1, x] - \mathbb{E}[f^*(S) \mid T = 0, x] \\
&= \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S)h(X',S) \mid S]}{\mathbb{E}_{X'}[w(X',S) \mid S]} \mid T = 1, \cancel{x}\right] - \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S)h(X',S) \mid S]}{\mathbb{E}_{X'}[w(X',S) \mid S]} \mid T = 0, \cancel{x}\right]
\end{aligned}$$

The conditioning on X in the outer expectations can be removed since $S \perp\!\!\!\perp X \mid T$ in that setting. Therefore, the learned $\tau_{f(S)}(x)$ is indeed constant w.r.t. x for setting b), but not necessarily equal to the true ATE.

Weights that yield the true ATE

Are there weights w that result in an unbiased estimate of the true ATE?

$$\begin{aligned}
\tau_{f^*(S)} &= \mathbb{E}[\tau_{f^*(S)}(X)] \\
&= \int_{x'} p(x') \int_s \frac{\int_x w(x,s)h(x,s)p(x \mid s)dx}{\int_x w(x,s)p(x \mid s)dx} \left(p(s \mid T = 1) - p(s \mid T = 0)\right) ds dx' \\
&= \int_s \frac{\int_x w(x,s)h(x,s)p(x \mid s)dx}{\int_x w(x,s)p(x \mid s)dx} \left(p(s \mid T = 1) - p(s \mid T = 0)\right) ds \\
&= \{w(x,s) = p(x)/p(x \mid s)\} \\
&= \int_s \frac{\int_x \frac{p(x)}{p(x|s)} h(x,s)p(x \mid s)dx}{\int_x \frac{p(x)}{p(x|s)} p(x \mid s)dx} \left(p(s \mid T = 1) - p(s \mid T = 0)\right) ds \\
&= \int_s \int_x p(x)h(x,s)dx \left(p(s \mid T = 1) - p(s \mid T = 0)\right) ds \\
&= \mathbb{E}_X[\mathbb{E}_S[h(X,S) \mid T = 1]] - \mathbb{E}_X[\mathbb{E}_S[h(X,S) \mid T = 0]] \\
&= \tau_Y .
\end{aligned}$$

Thus, for the choice of $w(x,s) = \frac{p(x)}{p(x|s)}$ the weighted regression will be an unbiased estimator of the ATE. But this is not the only choice.

Consider the weights $w^2(x,s) = (\frac{p(T=1|s,x)}{p(T=1|x)} - \frac{p(T=0|s,x)}{p(T=0|x)})^2$ and $w^+(x,s) = \frac{p(T=1|s,x)}{p(T=1|x)} + \frac{p(T=0|s,x)}{p(T=0|x)}$ for the CATE upper bound, described after Proposition 2. We ask what happens to the ATE estimate for the choice of $w^+(x,s)$. We begin noting that:

$$w^2(x,s) \underset{(a)}{=} \left(\frac{p(s|T=1,x)}{p(s|x)} - \frac{p(s|T=0,x)}{p(s|x)}\right)^2 \underset{(b)}{=} \frac{(p(s|T=0) - p(s|T=1))^2}{p(s|x)^2},$$

$$w^+(x,s) \underset{(a)}{=} \frac{p(s|T=1,x)}{p(s|x)} + \frac{p(s|T=0,x)}{p(s|x)} \underset{(b)}{=} \frac{p(s|T=0) + p(s|T=1)}{p(s|x)},$$

where (a) and (b) follows from Bayes' rule and $X \perp\!\!\!\perp S \mid T$ respectively. Looking at the ATE for f with w^+:

$$\begin{aligned}
\tau_{f(S)} &= \int_s \frac{\int_x w^+(x,s)h(x,s)p(x \mid s)dx}{\int_x w^+(x,s)p(x \mid s)dx} \left(p(s \mid T=1) - p(s \mid T=0)\right) ds \\
&\underset{(c)}{=} \int_s \frac{\int_x \frac{p(s|T=0)+p(s|T=1)}{p(s|x)} h(x,s)p(x \mid s)dx}{\int_x \frac{p(s|T=0)+p(s|T=1)}{p(s|x)} p(x \mid s)dx} \left(p(s \mid T=1) - p(s \mid T=0)\right) ds \\
&\underset{(d)}{=} \int_s \frac{\int_x \frac{p(x)}{p(s)} h(x,s)dx}{\int_x \frac{p(x)}{p(s)} dx} \left(p(s \mid T=1) - p(s \mid T=0)\right) ds \\
&= \int_s \int_x p(x)h(x,s)dx \left(p(s \mid T=1) - p(s \mid T=0)\right) ds \\
&= \mathbb{E}_X[\mathbb{E}_S[h(X,S) \mid T=1]] - \mathbb{E}_X[\mathbb{E}_S[h(X,S) \mid T=0]] \\
&= \tau_Y,
\end{aligned}$$

where the terms that depend only on s cancel in (c) and (d). Thus, in this case, the sum weights, w^+ are an unbiased estimator of the ATE. The same is true for any weighting function that can be written as $w(x,s) = \phi(s,t)/p(s \mid x)$, such as the absolute difference weights $w^1(x,s) = |\frac{p(T=1|s,x)}{p(T=1|x)} - \frac{p(T=0|s,x)}{p(T=0|x)}|$

Following a similar derivation for w^2, we get

$$\tau_{f(S)} = \int_s \frac{\int_x \frac{p(x)}{p(s|x)} h(x,s)dx}{\int_x \frac{p(x)}{p(s|x)} dx} \left(p(s \mid T=1) - p(s \mid T=0)\right) ds$$

which is not equal to τ_Y in general. However, in a randomized experiment version of case b), $S \perp\!\!\!\perp X$ and $p(s \mid x) = p(s)$. In this case, $\tau_{f(S)} = \tau_Y$. We leave the task of characterizing the precise bias in the general case for future work.

C.3 C) CONFOUNDED SURROGATE, NO EFFECT

In this case,

$$\mu_Y(x,t) = \mathbb{E}[Y(t) \mid X=x] = \mathbb{E}_{S(t)}[\mathbb{E}[Y \mid S(t), X=x] \mid X=x] = \mathbb{E}_S[\mathbb{E}[Y \mid X=x] \mid X=x, T=t] = \mathbb{E}[Y \mid X=x]$$

since $Y \perp\!\!\!\perp T \mid X$ and $Y \perp\!\!\!\perp S \mid X$. Since the potential outcomes are independent of T, with $h(x) = \mathbb{E}[Y \mid X=x]$

$$\tau_Y(x) = h(x) - h(x) = 0.$$

Trivially, there exists an $f(S)$ such that $\tau_Y(x) = \tau_{f(S)}(x)$ for all x, and that is any constant $f(S) = c$. Thus, $f = g$ is a minimizer of (2).

Since $Y \perp\!\!\!\perp S \mid X$, $h(x,s) = h(x)$. We can find the minimizer f^* of (2) by setting the derivative of any weighted regression to zero, conditioned on s,

$$\mathbb{E}[2w(X,s)(h(X) - f(s)) \mid S=s] = 0 \Longrightarrow f^*(s) = \frac{\mathbb{E}_X[w(X,s)h(X) \mid S=s]}{\mathbb{E}_X[w(X,s) \mid S=s]}.$$

This expression is not constant with respect to s in general. Then

$$\begin{aligned}
\tau_{f^*(S)}(x) &= \mathbb{E}[f^*(S) \mid T=1, x] - \mathbb{E}[f^*(S) \mid T=0, x] \\
&= \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',s)h(X') \mid S]}{\mathbb{E}_{X'}[w(X',s) \mid S]} \mid T=1, X=x\right] - \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',s)h(X') \mid S]}{\mathbb{E}_{X'}[w(X',s) \mid S]} \mid T=0, X=x\right] \\
&\neq 0
\end{aligned}$$

which is biased in general.

C.4 D) CONFOUNDED SURROGATE AND OUTCOME

In this case, the potential outcomes for the outcome $Y(1), Y(0)$ and surrogate $S(1), S(0)$ are independent given X,

$$\mu_Y(x,t) = \mathbb{E}[Y(t)|X=x] = \mathbb{E}_{S(t)}[\mathbb{E}[Y|S(t), X=x]|X=x] \underset{(a)}{=} \mathbb{E}_S[\mathbb{E}[Y|S,x]|X=x, T=t]$$

where in (a) we use $S(1), S(0) \perp\!\!\!\perp T|X$ and $Y \perp\!\!\!\perp T|X, S$. The treatment effect for $h(x,s) = \mathbb{E}[Y|X=x, S=s]$ is

$$\tau_Y(x) = \mathbb{E}_S[h(x,S) \mid T=1, X=x] - \mathbb{E}_S[h(x,S) \mid T=0, X=x]$$

which is non-zero in general, but constant when h is linear with respect to x, in that case $\tau_Y(x)$ is a constant. Since X carries more information on Y directly, than through S it is not possible to learn an f that minimizes (2). We can find the minimizer f^* by setting the derivative of any weighted regression to zero, conditioned on s,

$$\mathbb{E}_X[2w(X,s)(h(X,s) - f(s)) \mid S=s] = 0 \implies f^*(s) = \frac{\mathbb{E}_X[w(X,s)h(X,s) \mid S=s]}{\mathbb{E}_X[w(X,s) \mid S=s]}.$$

This expression is not constant with respect to s in general. Then

$$\begin{aligned}
\tau_{f^*(S)}(x) &= \mathbb{E}[f^*(S) \mid T=1, x] - \mathbb{E}[f^*(S) \mid T=0, x] \\
&= \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S)h(X',S) \mid S]}{\mathbb{E}_{X'}[w(X',S) \mid S]} \mid T=1, X=x\right] - \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S)h(X',S) \mid S]}{\mathbb{E}_{X'}[w(X',S) \mid S]} \mid T=0, X=x\right] \\
&\neq 0.
\end{aligned}$$

Unbiased weights Before we check if a choice of w results in true ATE, we note that the following identity follow from Bayes' rule

$$w^-(x,s) := \frac{p(T=1|x,s)}{p(T=1|x)} - \frac{p(T=0|x,s)}{p(T=0|x)} = \frac{\pi(x,s)}{p(s|x)}, \tag{16}$$

where $\pi(x,s) = p(s \mid x, T=1) - p(s \mid x, T=0)$, is the difference of densities. Now, we check for the weight w that result in the ATE:

$$\begin{aligned}
\tau_Y - \tau_{f^*(S)} &= \mathbb{E}_X[\mathbb{E}_S[\tau_{f^*(S)}|X=x]] - \mathbb{E}_X[\tau_Y(x)] \\
&= \mathbb{E}_X\left[\mathbb{E}_S[\tau_{f^*(S)}|X=x] - (\mathbb{E}_S[h(x,S) \mid T=1, X=x] - \mathbb{E}_S[h(x,S) \mid T=0, X=x])\right] \\
&= \int_x p(x) \int_s (f(s) - h(x,s))\, \pi(x,s) ds dx \\
&= \int_s \int_x (f(s) - h(x,s)) \underbrace{p(x)\pi(x,s)}_{Q(x,s)} dx ds \\
&= \int_s f(s) \int_x Q(x,s) dx ds - \int_s \int_x h(x,s) Q(x,s) dx ds \\
&= \int_s \frac{\int_{x'} w(x',s)h(x',s)p(x' \mid s)dx'}{\int_{x'} w(x',s)p(x' \mid s)dx'} \int_x Q(x,s)dxds - \int_s \int_x h(x,s)Q(x,s)dxds \\
&\text{Take } w = \frac{p(T=1|x,s)}{p(T=1|x)} - \frac{p(T=0|x,s)}{p(T=0|x)} \text{ and using (16):} \\
&= \int_s \frac{\int_{x'} \frac{Q(x',s)}{p(s)p(x'|s)} h(x',s)p(x' \mid s)dx'}{\int_{x'} \frac{Q(x',s)}{p(s)p(x'|s)} p(x' \mid s)dx'} \int_x Q(x,s)dxds - \int_s \int_x h(x,s)Q(x,s)dxds \\
&= \int_s \frac{\int_{x'} Q(x',s)h(x',s)dx'}{\int_{x'} Q(x',s)dx'} \int_x Q(x,s)dxds - \int_s \int_x h(x,s)Q(x,s)dxds \\
&= \int_s \int_{x'} Q(x',s)h(x',s)dx' - \int_s \int_x Q(x,s)h(x,s)dx \\
&= 0.
\end{aligned}$$

However, when $\frac{p(T=1|x,s)}{p(T=1|x)} < 1$. the weights $w^-(x,s)$ can be negative, leading to arbitrarily poor weighted regressions, and biased CATE in general.

C.5 E) DIRECT, UNMEDIATED EFFECT

In the case of a direct effect of T on Y, not mediated by S,

$\mu_Y(x,t) = \mathbb{E}[Y(t) \mid X = x] = \mathbb{E}_{S(t)}[\mathbb{E}[Y \mid S(t), X = x, T = t] \mid X = x] = \mathbb{E}_S[\mathbb{E}[Y \mid S, X = x, T = t] \mid X = x, T = t]$

and with a t dependent $h_t(x,s) = \mathbb{E}[Y \mid X = x, S = s, T = t]$,

$$\tau_Y(x) = \mathbb{E}_S[h_1(x,S) \mid T = 1, x] - \mathbb{E}_S[h_0(x,S) \mid T = 0, x]$$

There may not exist $f(S)$ such that $\tau_Y(x) = \tau_{f(S)}(x)$ for all x since X carries more information about Y. Since h depends on t, we can find the minimizer of f_t^* by setting the derivative of any weighted regression to zero, conditioned on s and t, which yields

$$f_t^*(s) = \frac{\mathbb{E}_X[w(X,s,t)h_t(X,s) \mid S = s, T = t]}{\mathbb{E}_X[w(X,s,t) \mid S = s, T = t]}$$

and gives the CATE estimate

$$\begin{aligned}\tau_{f^*(S)}(x) = \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S,t)h_1(X',S) \mid T = 1, S]}{\mathbb{E}_{X'}[w(X',S,t) \mid T = 1, S]} \mid T = 1, x\right] \\ - \mathbb{E}_S\left[\frac{\mathbb{E}_{X'}[w(X',S,t)h_0(X',S) \mid T = 0, S]}{\mathbb{E}_{X'}[w(X',S,t) \mid T = 0, S]} \mid T = 0, x\right]\end{aligned}$$

We can check what w results in true ATE,

$$\begin{aligned}\tau_{f^*(S)} &= \mathbb{E}_X[\mathbb{E}_S[\tau_{f^*(S)}|X = x]] = \int_x p(x) \int_s f(s)\pi(x,s)dsdx \\ &= \int_s f(s) \int_x p(x)\pi(x,s)dxds \\ &= \int_s \frac{\int_{x'} w(x',s,t)h_t(x',s)p(x' \mid s)dx'}{\int_{x'} w(x',s,t)p(x' \mid s)dx'} \int_x p(x)\pi(x,s)dxds \\ &\text{Take } w(x,s,t) = \frac{\pi(x,s)}{p(s|x)}, \text{ and simplify} \\ &\implies \int_s \frac{\int_{x'} h_t(x',s)p(x')\pi(x',s)dx'}{\int_{x'} p(x')\pi(x',s)dx'} \int_x p(x)\pi(x,s)dxds \\ &= \int_{x'} \int_s h_t(x',s)p(x')\pi(x',s)dx' \\ &= \mathbb{E}_X[\mathbb{E}_S[h_1(x,S)|X = x, T = 1]] - \mathbb{E}_X[\mathbb{E}_S[h_0(x,S)|X = x, T = 0]] = \tau_Y.\end{aligned}$$

However, if we are able to accurately estimate $h_1(x,s)$ and $h_0(x,s)$ ahead of the trial, there is little need for the trial in the first place, since their *averages* are precisely the goal of most experiments.

D EXPERIMENTAL DETAILS

D.1 SIMULATION STUDIES

We generate data starting using a binary treatment T, a vector-valued covariate X, and outcome Y. We generate three types of surrogate variables, S_{med}, S_{leaf}, S_{proxy}, to study how different baseline models learn about surrogates. S_{med} variables are mediators of the effect between T and Y, S_{leaf} variables are influenced by T but do not have an effect on Y, while the S_{proxy} variables influence Y but are not influenced by the treatment T. The structural equations are given as

$$\begin{aligned} X &\sim \mathcal{N}(\mu_x, \Sigma_X) \\ T &= \text{Bernoulli}\left(\sigma(W_{X\to T}^\top X + b_t + \epsilon_T)\right) \\ S_{\text{med}} &= b_{s_0} + W_{x\to S_0}^\top X + \left(W_{X\to S_0}^\top X + b_{s_0}\right) T + \epsilon_{S_0} \\ S_{\text{leaf}} &= b_{s_1} + W_{x\to S_1}^\top X + \left(W_{X\to S_1}^\top X + b_{s_1}\right) T + \epsilon_{S_1} \\ S_{\text{proxy}} &= b_{s_2} + W_{x\to S_2}^\top X + \epsilon_{S_2} \\ Y &= b_y + W_{X\to Y}^\top X + \phi(S_{\text{med}}) + \phi(S_{\text{proxy}}) + \epsilon_Y, \end{aligned} \tag{17}$$

where σ denotes the sigmoid function, and ϕ is a linear or a nonlinear map, depending on the scenario. We set X to have $k \in \{2, 5\}$ dimensions with standard normal distribution. The treatment parameters are $W_{X\to T} = [0.8, -0.6]$ with intercept term $b_t = -0.1$. Surrogate variables have a dimensionality of 3, 2, 2 for S_{med}, S_{leaf}, and S_{proxy} respectively. We choose standard normal distributions for the noise variables ϵ_T ϵ_{S_0}, ϵ_{S_1}, ϵ_{S_2}, ϵ_Y and the columns of the linear maps $W_{X\to S_0}$, $W_{X\to S_1}$, $W_{X\to S_2}$ are sampled from a uniform distribution on a hyper sphere of radius 0.7, 0.5, and 0.6 respectively. The intercept terms b_{s_0}, b_{s_1}, b_{s_2}, b_y are sampled from $\mathcal{N}(0.6, 0.25)$. Each experiment contains $N = 10000$ samples. We validate on a simulated trial data where we remove the confounding effect of X on T and sample $T \sim \text{Bernoulli}(0.5)$.

We modify the structural equations under scenarios (a)–(d) given in Figure 3—for instance by setting $W_{X\to Y} = 0$ for scenario (a). Furthermore each scenario has ten sub-scenarios, allowing us to study the baseline models under unobserved confounders and differing scales of relationships between the surrogates. In some scenarios, we add an unobserved confounder variable $U \sim \mathcal{N}(0, 1)$ influencing T, S, and Y, in other scenarios we change the relative scales of S_{med} and S_{proxy} and add non-linear square terms, $\phi(z) = z^2$, for their effects on the outcome. In total, this yields 60 experimental configurations[6].

D.2 METHODS

For **Surrogate sampling**, we first fit nuisance random forest models for $\hat{\mathbb{E}}[S|x, T = t]$ and sample $L = 50$ residuals of the fitted model via bootstrapping to simulate the conditional $\hat{p}(S|x, T = t)$. We fit a PyTorch (Paszke et al., 2019) linear regression model on the potential outcomes $\hat{s}(0), \hat{s}(1)$ sampled from the conditional distribution (see Algorithm 1). In our implementation, we use an unregularized linear specification (no L_1 penalty). For both methods, we estimate the surrogate index $\hat{h}(x, s)$ by fitting a gradient boost model. All models are implemented using scikit-learn implementation (Pedregosa et al., 2011). Random forest is fitted using default parameters, the parameters for gradient boost model is given in Table 4.

Density estimation Our bootstrapping approximation for simulating the conditional surrogate distribution is based on assumptions which hold for our simulated setting. We list the assumptions here for clarity and briefly discuss the implications.

1. The approach treats surrogates as conditionally independent given the context X and T,
$$p(S|X, T = t) = \prod_i p(S_i|X, T = t).$$
2. The random forest models estimate $\mathbb{E}[S_i|X, T = t]$ for all surrogate variables S_i
3. The noise in the synthetic setting is homoscedastic and mean-zero. Provided that the conditional mean is well fit by the random forest, the residuals for a given surrogate follow the same distribution marginally.

[6]Code and instructions for reproducing the experiments are available at https://github.com/Healthy-AI/causal-plugin-surrogates.

Table 4: Hyperparameters for the Surrogate Index Gradient Boost Model

Hyperparameter	Default Value
max_depth	6
min_samples_leaf	50
learning_rate	0.05
max_iter	400
l2_regularization	0.0
early_stopping	True
validation_fraction	0.1
n_iter_no_change	20

The strongest assumption here is that all of the surrogate variables are conditionally independent. In case this assumption does not hold, we could use a Markov factorization of surrogates instead, modeling each S_i as a function of *all* of its causes.

Bound regression We implement **Bound regression** by fitting a weighted regression of $\hat{h}$ on the surrogate vector S, where observations are weighted by the stabilized w^2, with propensities clipped between $(0.3, 0.7)$. In the synthetic experiments, we use a weighted Lasso implementation to encourage sparse, stable surrogate models. Propensity scores $e(x)$ and surrogate scores $\rho(x, s)$ are estimated using L_2-regularized logistic regression.

For the IHDP application, we instead use a weighted ordinary least squares (OLS) specification with sample weights, motivated by the more limited sample size and the resulting instability of regularization-based model selection in that setting. We use the fixed propensity specification, setting $e(X) = \hat{p}(T = 1)$, the empirical treatment probability in the training split. Since IHDP is a randomized experiment, treatment is independent of X by design, ensuring the constant propensity is correctly specified in this setting.

Optimal-tree variant In addition for **Bound regression** we implemented optimal tree-based model variant where we replace the linear lasso regression with an optimal tree model (Aglin et al., 2021). Since DL8.5 operates on binary input, we first discretize each surrogate S_j into a set of cumulative threshold indicators based on empirical quantiles.

$$Z_{j\ell} = \mathbb{I}\{S_j \le q_{j\ell}\}, \qquad q_{j\ell} \in \mathcal{Q},$$

and learn the tree on the resulting binary feature vector Z. Given predictions $\hat{h}_i = \hat{h}(X_i, S_i)$ and nonnegative weights w_i (defined by the chosen bound-weighting scheme), DL8.5 minimizes the squared error within the leaf. For a leaf L, the optimal leaf prediction is the weighted mean

$$\widehat{\mu}(L) \;=\; \frac{\sum_{i \in L} w_i \, \hat{h}_i}{\sum_{i \in L} w_i},$$

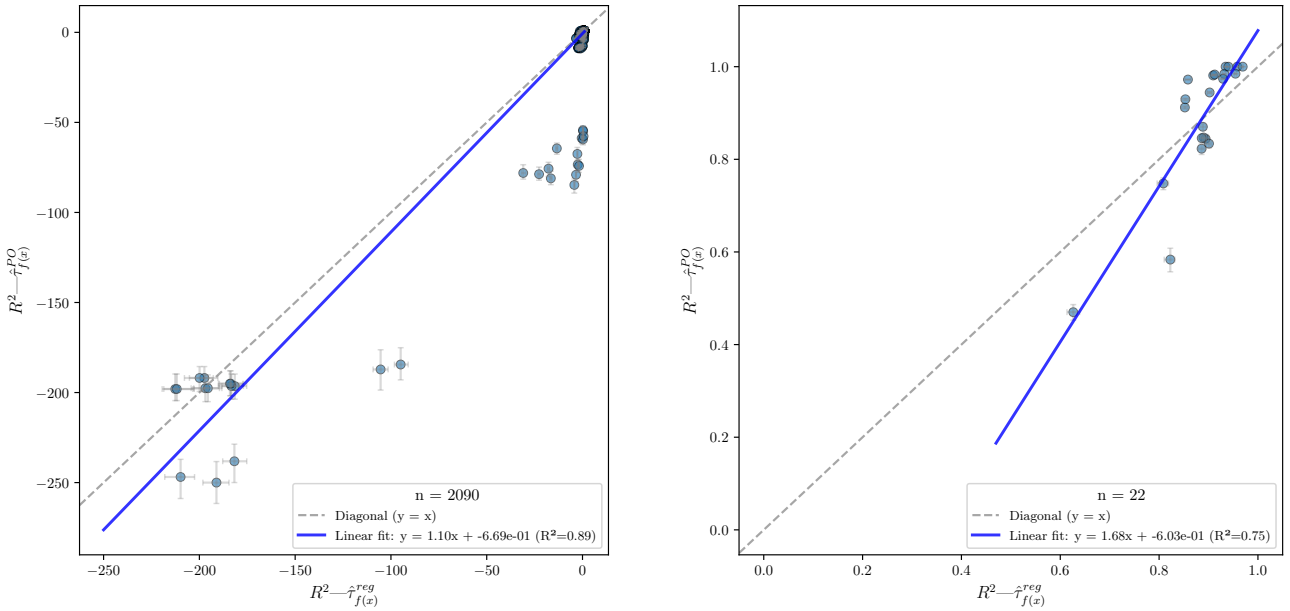
and the corresponding leaf loss is

$$\mathrm{Err}(L) \;=\; \sum_{i \in L} w_i \big(\hat{h}_i - \widehat{\mu}(L)\big)^2.$$

The tree objective is the sum of $\mathrm{Err}(L)$ over leaves, subject to the depth, minimum-support, and run time constraints of DL8.5.

Baseline models We include outcome regressions, estimating $\mathbb{E}[Y \mid S = s]$, as well a single-surrogate regression (Reg-Sel-Reg), learned by fitting a regression, selecting the variable with the largest absolute coefficient, and regressing only on that. We also include the surrogate index $h(x, s) = \hat{\mathbb{E}}[Y|x, s]$ as a comparison that depends directly on pre-treatment variables, x. All baselines come in linear and tree model variants, with an additional gradient boosting regression with parameters in Table 4 for the surrogate index.

Hyperparameter search for tree models For the tree models, we use a $5-$fold cross validation and use grid search over hyperparameters consisting of tree depth, necessary numbers for splitting, and number of minimum samples per leaf, and the complexity regularization parameter. See Table 5 for the grid search parameters.



(a) Composite scenarios in Table 1

(b) Linear scenarios in Table 10

Figure 4: Comparison of R^2 computed via potential outcomes and regression across reported results. We omit results from scenarios b) and c) as the CATE is constant and R^2 is not meaningful.

Table 5: Grid search parameter space for the tree-based regression models. The grid controls tree depth, minimum leaf size, minimum split size, and cost-complexity pruning strength.

Parameter	**Values**
`reg__max_depth`	[3, 4, None]
`reg__min_samples_leaf`	[50, 100, 200]
`reg__min_samples_split`	[100, 200, 400]
`reg__ccp_alpha`	[0.0, 10^{-3}]

D.3 METRICS

We report mean absolute error (MAE $\hat{\tau}$), for the treatment effect τ ATE, and precision in average treatment effect (PEHE $= \frac{1}{N}\sum_{i=1}^{N}\ (\tau(x_i) - \hat{\tau}(x_i))^2$) and coefficient of determination R^2 for the CATE, $\tau(x)$. We estimate $\hat{\tau}$ using the estimated average potential outcomes in the trial $\hat{\tau} = \hat{\mathbb{E}}^e[f(S) \mid T = 1] - \hat{\mathbb{E}}^e[f(S) \mid T = 0]$. However, such averaging techniques are not feasible for estimating CATE as both potential outcomes $s(1)$, $s(0)$ are not available. We estimate CATE using two methods, first by fitting linear models $g_t(x)$ for $\mathbb{E}_S[f(S)|x, T = t]$, and by using the ground-truth potential outcomes $s(1)$ and $s(0)$ and then estimate the CATE using

$$\hat{\tau}^{reg}_{f(S)}(x) = g_1(x) - g_0(x),$$
$$\hat{\tau}^{PO}_{f(S)}(x) = f(s(1)) - f(s(0)), \text{ where } s(t) \sim p(s|x, T = t),$$

for each method, respectively. In Table 1 and Table 10, we report the regression based metric as the R^2. We give a comparison of the cases in Figure 4.

D.4 IHDP VARIABLES

For lists of the variables used in IHDP, see Tables 6–9.

Table 6: Design and treatment variables in IHDP.

Variable	Role	Description
IHDP_Number	ID	Infant identifier.
Site_name	Site	Multi-site trial identifier.
Treatment_group	T	Randomized assignment (mapped to 0/1).
Birth_weight_group	Stratum	Birth-weight group (used for subgroup analysis).

Table 7: Post-treatment surrogate variables S (SURROGATE_SET=all).

Time	Variables
40 weeks	Infant_weight_40w, Infant_length_40w, BMI_40w
4 months	Infant_weight_4m, Infant_length_4m, BMI_4m
8 months	Infant_weight_8m, Infant_length_8m, BMI_8m
12 months	Anthropometrics + Bayley_MDI_12m, Bayley_PDI_12m, Stein/RAND 12m scales
18 months	Infant_weight_18m, Infant_length_18m, BMI_18m
24 months	Anthropometrics + Bayley_MDI_24m, Bayley_PDI_24m, Stein/RAND 24m scales
Year 1–2 summaries	Morbidity indices, serious conditions, hospitalizations, injuries, illnesses (_Y1, _Y2)

Table 8: Baseline covariates X (pre-treatment; suffix _baseline).

Variable	Description
Birth_weight_gm_baseline	Birth weight (grams).
Infant_sex_baseline	Sex (1=M, 2=F).
Maternal_age_birth_baseline	Maternal age at birth (years).
Maternal_education_baseline	Education (1:9th; 2:9–12th; 3:HS; 4:some college; 5:college+).
Maternal_race, BLACK, HISPANIC	Race/ethnicity indicators.
Neonatal_health_index_by_BW_baseline	Neonatal health index.
Analysis_gestational_age_weeks_baseline	Gestational age (weeks).
SGA_based_on_ANGA_BW_baseline	Small-for-gestational-age (0/1).
SGA_based_on_ANGA_birth_length_baseline	SGA by birth length (0/1).
SGA_based_on_ANGA_birth_head_circ_baseline	SGA by head circumference (0/1).
Birth_order_baseline	Birth order (1,2,3+).
Marital_status_birth_baseline	Marital status (1–4).
Head_circ_birth_cm_baseline	Head circumference at birth (cm).
Infant_length_at_birth_baseline	Length at birth.
BMI_at_birth_baseline	Body mass index at birth.

Table 9: Primary outcome.

Variable	Time	Description
Stanford_binet_IQ_36m	36 months	Stanford–Binet IQ score.

E ADDITIONAL EXPERIMENTAL RESULTS

In Table 10, we show the results for all methods on linear versions of scenarios a–d).

Table 10: Results on linear synthetic scenarios a–d), matching the graphs in Figure 1 and described further in Appendix D. Results are averaged over linear and nonlinear scenarios. We report 95% confidence intervals computed using the standard error for each run. †The surrogate index does not produce a plug-in surrogate as it depends on pre-treatment variables.

Linear Scenarios	**Case a)** $\overline{ATE}$=1.83		**Case b)** $\overline{ATE}$=2.11	**Case c)** $\overline{ATE}$=0.00	**Case d)** $\overline{ATE}$=18.54	
Method	MAE $\hat{\tau}\downarrow$	$R^2\hat{\tau}(x)\uparrow$	MAE $\hat{\tau}\downarrow$	MAE $\hat{\tau}\downarrow$	MAE $\hat{\tau}\downarrow$	$R^2\hat{\tau}(x)\uparrow$
Reg.-sel.-reg. (linear)	0.05± 0.04	0.86± 0.01	0.73± 0.02	2.45± 0.03	1.98± 0.06	0.85± 0.01
Reg.-sel.-reg. (tree)	0.06± 0.05	0.85± 0.01	0.74± 0.02	2.32± 0.03	1.40± 0.08	0.81± 0.01
Outcome reg. (linear)	0.16± 0.04	0.91± 0.01	0.03± 0.03	2.03± 0.03	1.63± 0.05	0.90± 0.00
Outcome reg. (tree)	0.14± 0.05	0.89± 0.01	0.05± 0.03	1.03± 0.05	0.93± 0.07	0.89± 0.01
Surrogate index† (linear)	0.02± 0.03	0.93± 0.00	0.02± 0.03	0.01± 0.01	0.04± 0.04	0.96± 0.00
Surrogate index† (tree)	0.08± 0.04	0.89± 0.01	0.22± 0.03	0.72± 0.04	0.16± 0.08	0.90± 0.01
Surrogate index† (hgb)	0.03± 0.03	0.93± 0.00	0.02± 0.02	0.14± 0.02	0.07± 0.05	0.95± 0.00
Bound reg. (linear)	0.14± 0.05	0.91± 0.00	0.10± 0.03	1.48± 0.03	1.11± 0.05	0.93± 0.00
Bound reg. (tree)	0.13± 0.05	0.89± 0.01	0.01± 0.02	0.43± 0.04	1.27± 0.06	0.89± 0.01
Surrogate Sampl. (linear)	0.03± 0.03	0.94± 0.00	0.01± 0.02	0.09± 0.01	0.03± 0.03	0.97± 0.00

F LIMITATIONS OF EXISTING AND ALTERNATIVE METHODS

F.1 COMPARISON TO ATHEY ET AL.

Our work shares a similar setting to work of Athey et al. (2025) in which the treatment-outcome (T, Y) pairs are not observed together. We also make use of the "surrogates index" h for learning the composite plug-in surrogate $f(S)$. However, we differ in our goals, our learning objective, our methods, and experiments. We give a non-exhaustive comparison of our differences.

1. **Goals** The plug-in surrogate must depend only on post-treatment variables and be interpreted as a surrogate outcome itself D3. This does not apply to the surrogate index (Athey et al., 2025).
2. **Learning objective** Our goal demands an entirely different learning objective (2), which we define as matching the in-trial conditional average treatment effect (CATE). Athey et al. (2025) focus instead on an unbiased estimator of the in-trial ATE. Our strategies overlap in the use of the nuisance function $h(x, s)$ (1), but deviate immediately after. We show that finding an exact plug-in surrogate is not always possible, but give conditions under which it is, along with recommendations for how to achieve it.
3. **Model classes** Our learning objective is designed for learning small-model surrogates, such as trees or linear models, as used in manual composite surrogates. Fitting $h(x, s)$ using such a model may fail to account for confounders, or may lose dependence on s if $\mathbb{E}[Y \mid x, s]$ varies mostly with x.
4. **Algorithms** Our surrogate sampling and bound regression algorithms directly target minimization of our proposed learning objective and are new.
5. **Experiments** Our empirical study uses entirely new synthetic tasks and real-world data.

F.2 ADDITIONAL RELATED WORK

As several definitions of surrogates rely on untestable (causal) assumptions, researchers have studied the predictiveness of various surrogates empirically (Gilbert and Hudgens, 2006; Elliott et al., 2015; Wang et al., 2022). Recently, a study examining this question using surrogate outcomes in A/B tests at a large streaming service (Zhang et al., 2023) found that the so-called "surrogate index" (Athey et al., 2025), based on the definition of Prentice (1989), showed strong predictive

performance. In a recent work Molenberghs et al. (2024) suggest a causally grounded method of adopting and evaluation of surrogates in clinical trials.

F.3 OUTCOME REGRESSION IS CONFOUNDED IN GENERAL

Regression of Y onto S to select a surrogate yields confounded estimates of the treatment effect in general. In Proposition 3 we show how such confounding bias emerges, and give a simple example scenario to highlight the bias.

Proposition 3 (Biased Linear Estimator)**.** *Suppose that the outcome is given by $Y(t) = \gamma^\top X + \delta^\top S(t) + \epsilon_Y$, with noise variable ϵ_Y. A linear estimator for $\mathbb{E}[Y|S]$, will be a biased estimator of the treatment effect τ_Y.*

Proof. A linear model recovers $\mathbb{E}[Y|S] = \gamma^\top \mathbb{E}[X|S] + \delta^\top S$, where $E[X|S]$ is not necessarily linear. Resulting in a bias in the estimated effect

$$\begin{aligned}\mathbb{E}[Y|S(1)] - \mathbb{E}[Y|S(0)] &= \gamma^\top(\mathbb{E}[X|S(1)] - \mathbb{E}[X|S(0)]) + \delta^\top \tau_S \\ &\neq \delta^\top \tau_S = Y(1) - Y(0).\end{aligned}$$

□

Example Scenario Consider the following structural equations:

$$\begin{aligned}X &\sim \mathcal{N}(0,1) \\ T &\sim \text{Bernoulli}(0.5) \\ S &\sim \mathcal{N}(XT + T, 1) \\ Y &\sim \mathcal{N}(5X + S, 1)\end{aligned}$$

In this case, it is easy to prove that $\tau_Y = \mathbb{E}[Y(1) - Y(0)] = 1$. However, using $f(S) = \mathbb{E}[Y \mid S] = 5\mathbb{E}[X \mid S] + S$ as a plug-in surrogate,

$$\tau_{f(S)} = \mathbb{E}_X\left[\mathbb{E}_S[f(S) \mid X, T=1] - \mathbb{E}_S[f(S) \mid X, T=0]\right] \approx 2.43\,.$$

Note that this is not because of confounding of the effect of T on Y since T is marginally independent of X (and is causally unaffected by all other variables). This is merely because the association between S and Y is confounded by X.

F.4 MATCHING ATES $\mathbb{E}[\Delta_{f(S)}]$ AND $\mathbb{E}[\Delta_Y]$ IS NOT ENOUGH

We give a simple result showing why matching average treatment effects is a weak criterion for surrogate learning.

Theorem 1. *Assume that X is an adjustment set for the effect of T on S, that is $S(t) \perp\!\!\!\perp T \mid X$ and that $Y \perp\!\!\!\perp T \mid X, S$. Further, assume that, for some surrogate variable $S_i \in S$, there exists $\delta > 0$, such that $|\mathbb{E}[w_1(X)S_i \mid T=1] - \mathbb{E}[w_0(X)S_i \mid T=0]| \geq \delta$, where $w_t(x) = \frac{p(T=t)}{p(T=t|X=x)}$ for $t \in \{0,1\}$. Then,*

$$f(S) = \frac{\mathbb{E}[\Delta_Y]}{\delta} S_i \quad \textit{satisfies} \quad \mathbb{E}[\Delta_{f(S)}] = \mathbb{E}[\Delta_Y]\,.$$

Proof. We show later that

$$\begin{aligned}\mathbb{E}[Y(t)] &= \mathbb{E}_{X,S}\left[w_t(x)h(X,S) \mid T=1\right] \\ \mathbb{E}[f(S(t))] &= \mathbb{E}_{X,S}\left[w_t(x)f(S) \mid T=1\right].\end{aligned}$$

To match the ATEs of the surrogate to the primary outcome, we therefore need

$$\mathbb{E}\left[w_1(X)f(S) \mid T=1\right] - \mathbb{E}\left[w_0(X)f(S) \mid T=0\right] = \mathbb{E}[\Delta_Y]$$

As long as $\delta = \mathbb{E}[w_1(X)S_i \mid T=1] - \mathbb{E}[w_0(X)S_i \mid T=0]$ satisfies $|\delta| > 0$ for some surrogate variable S_i, this is trivially achievable by a linear $f(S) = \alpha S_i$ since then

$$\mathbb{E}\left[w_1(X)f(S) \mid T=1\right] - \mathbb{E}\left[w_0(X)f(S) \mid T=0\right] = \alpha\delta$$

and we can define $\alpha = \mathbb{E}[\Delta_Y]/\delta$. In other words, as long as the treatment T has an effect on *any* part of S, we can match the average causal effect with some model $f(S)$. □

Most choices of α, S_i, however, would not be good choices for a surrogate. For example, the statistical implications of choosing a surrogate barely affected by the treatment may be large—it would require a large experiment cohort to see an effect. To select α optimally requires knowing $\mathbb{E}[\Delta_Y]$ precisely.