Benchmarking Observational Studies with Experimental Data under Right-Censoring

Ilker Demirel
MIT

Edward De Brouwer
Yale University

Zeshan Hussain
MIT

Michael Oberst
Carnegie Mellon University

Anthony Philippakis
Broad Institute of MIT and Harvard

David Sontag
MIT

Abstract

Drawing causal inferences from observational studies (OS) requires unverifiable validity assumptions; however, one can *falsify* those assumptions by benchmarking the OS with experimental data from a randomized controlled trial (RCT). A major limitation of existing procedures is not accounting for *censoring*, despite the abundance of RCTs and OSes that report right-censored time-to-event outcomes. We consider two cases where censoring time (1) is independent of time-to-event and (2) depends on time-to-event the same way in OS and RCT. For the former, we adopt a censoring-doubly-robust signal for the conditional average treatment effect (CATE) to facilitate an equivalence test of CATEs in OS and RCT, which serves as a proxy for testing if the validity assumptions hold. For the latter, we show that the same test can still be used even though unbiased CATE estimation may not be possible. We verify the effectiveness of our censoring-aware tests via semi-synthetic experiments and analyze RCT and OS data from the Women's Health Initiative study.

1 INTRODUCTION

The ability to reliably establish causal relationships is essential for decision-making and policy development (Pearl and Mackenzie, 2018; Angrist and Pischke, 2009; Hernan and Robins, 2021). Although experimental data, often collected through randomized controlled trials (RCT), are considered to be the "gold-standard" for inferring causality, real-world evidence collected from observational (non-experimental) data is increasingly guiding regulatory processes (Hansford et al., 2023; Government of Canada, 2019; Therapeutic Goods Administration, 2023; NICE, 2022). Indeed, observational data, such as claims and electronic health records, can provide large-scale, diverse, and longitudinal data at low cost, making it a promising complement to time and cost-intensive experimental data.

Ideally, one would like to leverage observational studies (OS) when experimental data is unavailable or provides limited evidence (Dagan et al., 2021; Gershman et al., 2018). For instance, people with a history of cardiovascular diseases may not be eligible to participate in an RCT. Therefore, OSes are the only source of data for those people. Furthermore, the limited sample size of RCTs makes subgroup analysis infeasible and their results do not always apply to a *target* population of interest (Rothwell, 2005; Stuart et al., 2011; Hartman et al., 2015; Colnet et al., 2020). Contrary to the RCTs, however, OSes are susceptible to numerous sources of bias, bringing their utility in practice under question. Therefore, it is critical to evaluate the credibility of an OS before using it for different downstream tasks (Yang et al., 2023). To that end, we will develop a hypothesis test to check if the findings from an OS and an RCT are compatible within the trial-eligible population, that can be used when the outcomes are *right-censored* (Kalbfleisch and Prentice, 2011).

Naive analysis of an OS can lead to biased effect estimates due to various reasons. Among those, unobserved confounding— which makes prognostic factors systemically differ in treatment and control groups— typically receives the most attention. However, the bias may also emerge due to the poor analysis of the data regarding the handling of censored outcomes and

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

non-adherence to treatment assignments (Hernán et al., 2017), the definition of time-zero and follow-up (Lodi et al., 2019; Hernán et al., 2008), and different types of *selection bias* (Hernán et al., 2004; Yadav and Lewis, 2021). Target trial emulation (TTE), where one uses observational data to emulate a hypothetical trial, has emerged as a popular framework to limit the bias in the OSes (Hernán and Robins, 2016; Wang et al., 2023).

With a well-specified TTE protocol, it is possible to estimate causal effects from observational data under well-known *internal* and *external* validity assumptions (given in Section 2) (Imbens and Rubin, 2015; Wager and Athey, 2018; Semenova and Chernozhukov, 2021). Internal validity ensures that the causal effects can be reliably inferred in the *OS population*. External validity further allows transporting those effect estimates to different populations (*e.g.*, the RCT population) (Dahabreh et al., 2020a). Even though these assumptions are not *verifiable*, one can still *falsify* them by benchmarking the OS to an RCT (Dahabreh et al., 2020b; Forbes and Dahabreh, 2020). The key idea is to formulate and test a null hypothesis that captures the implications of those assumptions, which is the equivalence of the treatment effects inferred from the OS and the RCT. The rejection of the null would then be linked to the violation of (a subset of) those assumptions.

Recent works have developed tests for falsifying the internal and external validity assumptions. Hussain et al. (2022) proposed an algorithm to first compare the *group-level* effects derived from an RCT and multiple OSes in pre-specified groups and integrate the evidence only from the OSes compatible with the RCT. Hussain et al. (2023) developed a falsification framework that compares *individual-level* effect estimates from an RCT and an OS over the entire covariate space and *automatically* detects the regions of disparity, providing explanations in the form of witness functions. De Bartolomeis et al. (2023) adopt an alternative view and focus on quantifying the hidden confounding in an OS from a "sensitivity analysis" perspective instead of testing for the equivalences of the effect estimates across studies. Karlsson and Krijthe (2024) show how one can detect hidden confounding when there is no RCT data but *multiple* OSes that share a data-generating graph with certain properties. None of the studies above consider censored observations.

Our Contributions Censoring due to drop-outs or loss-to-follow-ups is a common issue that plagues both OSes and RCTs. Improperly handling the censored data does not merely lead to a suboptimal falsification test for the validity assumptions but renders the test unreliable since censoring can easily introduce bias. We generalize the test in Hussain et al. (2023) to cases

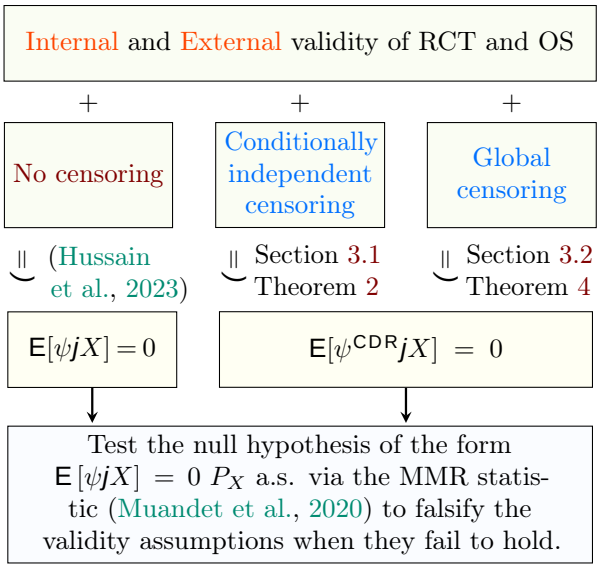

Figure 1: Prior work develops a maximum moment restriction (MMR)-based falsification test for validity assumptions, under no censoring. We extend the test to the case where time-to-event outcomes are right-censored, considering two censoring mechanisms.

with right-censored time-to-event outcomes. We first consider in Section 3.1 the common scenario where the censoring time is conditionally independent of the time-to-event. In Section 3.2, we introduce a novel censoring concept, *global censoring*, where the censoring time depends on time-to-event (*e.g.*, drop-out due to disease progression), but in the same way in the RCT and the OS. We develop a falsification test under both censoring mechanisms and verify its effectiveness in semi-synthetic experiments with the Infant Health and Development Program cohort. We also analyze real-world RCT and OS data from the Women's Health Initiative. Figure 1 gives an overview of our results.

2 NOTATION AND BACKGROUND

Notation Let $A \in \{0,1\}$ and $Y \in \mathbb{R}_+$ denote the binary treatment assignment and time-to-event outcome. We use $Y(a)$ to refer to the *potential outcome* for treatment $A = a$ and use S as the study indicator with $S = 0$ reserved for the RCT and $S = 1$ for the OS. We denote by $\mathcal{I}_0$ and $\mathcal{I}_1$ the set of patients for the RCT and the OS, respectively, and let $\mathcal{I} := \mathcal{I}_0 \cup \mathcal{I}_1$. We denote the cardinality of a set by $|\cdot|$ and let $|\mathcal{I}_0| = n_0$, $|\mathcal{I}_1| = n_1$, and $|\mathcal{I}| = n$, where $n = n_0 + n_1$.

We denote the set of patient covariates by X and define $\mathcal{X}$ as the space of trial-eligible patients, that is,

$$P(S = 0 \mid X = x) > 0, \qquad \forall x \in \mathcal{X}.$$

Internal and External Validity We first introduce conditional average treatment effect (CATE) estimation without censoring. For a patient with covariates $x \in \mathcal{X}$ in study $S = s$, the CATE is defined as the expected difference between the *potential outcomes* with and without treatment:

$$\text{CATE}(x, s) \coloneqq \mathbb{E}\left[Y(1) \quad Y(0) | X = x, S = s\right]. \quad (1)$$

The fundamental challenge in causal inference is that the potential outcomes $Y(0)$ and $Y(1)$ are never observed together. For a patient in the control group (*i.e.*, $A = 0$) we observe $Y(0)$ but not $Y(1)$, and vice versa for the treatment group. Nevertheless, one can still estimate the CATE in a given study $S = s$ under certain internal validity assumptions listed below.

Definition 1 (*Internal validity*)**.** *We say that internal validity holds for a study $S = s$ if the following conditions hold $\forall a \in \{0, 1\}$ and $\forall x \in \mathcal{X}$.*

- *No unobserved confounding* — $Y(a) \perp\!\!\!\perp A \mid X, S = s$.
- *Consistency* — $A = a \implies Y = Y(a)$.
- *Positivity of treatment assignment*

$$P\left(A = a \mid X = x, S = s\right) > 0.$$

Assumption 1. *Internal validity (Definition 1) holds for both the RCT $S = 0$ and the OS $S = 1$.*

Assumption 1 allows to identify the CATE in (1) as a quantity that can be estimated from data (*i.e.*, an estimand) in the RCT population and the OS population *separately*. However, even when the internal validity holds, the CATE in two populations can differ due to different distribution of treatment effect modifiers that are not included in X. To generalize the CATE from one study to the other, one needs external validity that assumes away unobserved treatment effect modifiers (Dahabreh et al., 2019, 2020a).

Assumption 2 (*External validity*)**.** *We have, $\forall a \in \{0, 1\}$ and $\forall x \in \mathcal{X}$,*

- *Ignorability of selection* — $Y(a) \perp\!\!\!\perp S \mid X$.
- *Positivity of selection* — $P\left(S = 1 \mid X = x\right) > 0$.

Falsification Without Censoring Assumption 1 implies that $\text{CATE}(x, 1)$ and $\text{CATE}(x, 0)$ can be estimated, and Assumption 2 implies that $\text{CATE}(x, 1) = \text{CATE}(x, 0)$. Intuitively, disagreement between the estimated CATE functions from each study implies that one or more of these assumptions are violated. Testing this equivalence forms the basis of the maximum moment restriction (MMR)-based falsification framework proposed in Hussain et al. (2023) in the *absence of censoring*. The core technical idea is to relate the equivalence of the underlying CATE functions to a set of conditional moment restrictions, finding an "instance-wise signal" ψ such that Assumptions 1-2 imply $\mathbb{E}\left[\psi \mid X\right] = 0$ almost surely. This reduction allows for applying the recent advances in the testing of MMRs via kernel methods (Muandet et al., 2020).

3 CENSORED FALSIFICATION

This work considers the common scenario where the time-to-event outcome $Y \in \mathbb{R}_+$ is subject to right-censoring. Let $C \in \mathbb{R}_+$ be the censoring time. For a patient, we observe $(X, A, S, \tilde{Y}, \Delta)$ where

$$\tilde{Y} = \min\left(Y, C\right), \qquad \Delta = \mathbf{1}\{Y \leq C\}. \quad (2)$$

We either observe the time-to-event or the censoring time, indicated by Δ. $\Delta = 1$ means that the time-to-event Y is observed, but not the censoring time C, and vice versa for $\Delta = 0$. Censoring introduces an additional identification problem as neither of the potential outcomes is observed for censored patients. Without any assumptions on the censoring mechanism, unbiased CATE estimation is not possible in either study, rendering prior falsification approaches ineffective.

In this section, we generalize the falsification framework in Hussain et al. (2023) under two different censoring conditions. In Section 3.1, we assume the censoring time C is independent of the time-to-event Y after conditioning on covariates X. We derive unbiased instance-wise signals for CATE in the RCT and the OS, suitable for comparison via an MMR test. Note that the falsification could be due to violating the censoring assumption, even though the validity assumptions hold. Remarkably, even when the censoring is not conditionally independent of time-to-event, and therefore unbiased CATE estimation is infeasible, testing the validity assumptions using the same signals may still be possible. The key ingredient is an alternative assumption that the censoring mechanism is identical across studies (Section 3.2).

3.1 Falsification with Conditionally Independent Censoring

We start by recalling the conditionally independent censoring condition, which is common in survival analysis (Kalbfleisch and Prentice, 2011).

Assumption 3 (*Conditionally independent censoring*)**.**

$$Y \perp\!\!\!\perp C \mid X, A, S.$$

We show that under conditionally independent censoring, one can adapt the doubly-robust censoring-unbiased estimator in Rubin and van der Laan (2007)

to form conditional moment restrictions (CMR) that enables us to use an MMR-based test (Muandet et al., 2020) for falsifying the validity assumptions (1-2). Precisely speaking, we derive instance-wise signals ψ_s such that $\mathsf{E}[\psi_s \mid X] = \text{CATE}(X, s)$, allowing us to define a difference signal $\psi = \psi_1 - \psi_0$ such that $\mathsf{E}[\psi \mid X] = 0$ almost surely in P_X. We can then use the preceding equality as our null hypothesis, whose rejection would imply the violation of the validity assumptions, and use an MMR-based test similar to that of Hussain et al. (2023) (see Theorem 3).

By G and F, we denote the cumulative distribution functions of the censoring time C and time-to-event outcome Y, respectively. We use $\bar{G}$ and $\bar{F}$ to denote their *survival* functions ($\bar{G}(t) = 1 - G(t)$). For conciseness, we use the following notation

$$\bar{G}_{s,a}(t \mid X) \coloneqq P(C > t \mid X, S = s, A = a). \quad (3)$$
$$\bar{F}_{s,a}(t \mid X) \coloneqq P(Y > t \mid X, S = s, A = a). \quad (4)$$

We further assume that for any realizable time-to-event, there is a nonzero probability of being observed, to prevent censoring-related identifiability issues.

Assumption 4 (*Support under censoring*). *$\forall t \in \mathsf{R}_+$, $\forall a, s \in \{0, 1\}$, the following holds almost surely in P_X*[1]

$$\bar{F}_{s,a}(t \mid X) > 0 \implies \bar{G}_{s,a}(t \mid X) > 0.$$

Combined with internal validity in Assumption 1, Assumptions 3 and 4 make it possible to have unbiased estimates of the CATE (see (1)) in the RCT and the OS populations in the presence of right-censoring.

We start by writing the censoring-unbiased signal from Rubin and van der Laan (2007) for a study s and treatment group a pair.

$$\psi_{s,a} = \frac{\mathbf{1}\{\Delta = 1\} Y}{\bar{G}_{s,a}(Y \mid X)} + \frac{\mathbf{1}\{\Delta = 0\} Q_{s,a}(X, C)}{\bar{G}_{s,a}(C \mid X)} - \int_1^{\tilde{Y}} \frac{Q_{s,a}(X, c)}{\bar{G}^2_{s,a}(c \mid X)} dG_{s,a}(c \mid X), \quad (5)$$

where $Q_{s,a}(X, C) = \mathsf{E}[Y \mid X, Y > C, S = s, A = a]$. $\psi_{s,a}$ is "doubly-robust" in the sense that it is unbiased for the time-to-event outcome Y if either $\bar{G}_{s,a}(t|X)$ or $\bar{F}_{s,a}(t|X)$ is correctly estimated. Computing $\psi_{s,a}$ requires estimating the survival function of the censoring time, $\bar{G}_{s,a}(t|X)$, and of the time-to-event outcome, $\bar{F}_{s,a}(t|X)$. Using the estimate for $\bar{F}_{s,a}(t|X)$, one can also calculate $Q_{s,a}(X, C)$ by integration under conditionally independent censoring. The practitioner may use covariate-adjusted Kaplan-Meier estimators or the Cox proportional hazards (CoxPH) framework to model the effect of covariates on the time-to-event along with recent advances in survival modeling (Van Keilegom et al., 2001; Cole and Hernán, 2004; Cox, 1972; Chapfuwa et al., 2021; Curth et al., 2021). We adopt a CoxPH model and provide the details in Section 4.

Lemma 1. *Suppose that Assumptions 1,3, and 4 hold. From Theorem 1 in Rubin and van der Laan (2007) and Assumption 1, we have, $\forall s, a \in \{0, 1\}$,*

$$\mathsf{E}[\psi_{s,a} \mid X, S = s, A = a] = \mathsf{E}[Y(a) \mid X, S = s],$$

if $\bar{G}_{s,a}(t|X)$ (3) or $\bar{F}_{s,a}(t|X)$ (4) is correctly estimated.

Lemma 1 is quite powerful as it identifies conditional average potential outcomes under right-censoring in doubly-robust way. However, $\psi_{s,a}$ is not readily available as an *instance-wise* signal to facilitate an MMR-based test, as it is only defined for $S = s$ and $A = a$. One can handle this by re-weighting with inverse selection $P(S|X)$ and propensity scores $P(A|X, S)$.

Corollary 1. *Suppose that Assumptions 1,3,4 hold and $P(S = 1|X)$ is correctly estimated. Let*

$$\psi^{\mathsf{IPW},}_{s,a} \coloneqq \frac{\mathbf{1}\{S = s, A = a\} \psi_{s,a}}{P(S = s \mid X) P(A = a \mid X, S = s)}. \quad (6)$$

Then $\forall s, a \in \{0, 1\}$, $\mathsf{E}[\psi^{\mathsf{IPW},}_{s,a} \mid X] = \mathsf{E}[Y(a)|X, S = s]$ if $P(A = a|X, S = s)$, and either $\bar{G}_{s,a}(t|X)$ (3) or $\bar{F}_{s,a}(t|X)$ (4) are correctly estimated.

Following Corollary 1, one can define the instance-wise signal $\psi^{\mathsf{IPW},}_s \coloneqq \psi^{\mathsf{IPW},}_{s,a=1} - \psi^{\mathsf{IPW},}_{s,a=0}$ which is unbiased for the CATE in study s, that is, $\mathsf{E}[\psi^{\mathsf{IPW},}_s \mid X] = \text{CATE}(X, s)$ (see (1)). We can then test the equivalence $\text{CATE}(X, 1) = \text{CATE}(X, 0)$ via an MMR test with the null hypothesis of $\mathsf{E}[\psi^{\mathsf{IPW},}_{s=1} - \psi^{\mathsf{IPW},}_{s=0} \mid X] = 0$.

Enhancing Double-robustness While we are now well-equipped for an MMR-based falsification test, Corollary 1 requires the correct estimation of the propensity score of treatment $P(A = a|X, S = s)$. We alleviate this requirement by building upon the doubly-robust estimation of treatment effects literature, where the correct estimation of either the propensity score or the mean outcome function is sufficient. In particular, since $\bar{F}_{s,a}(t|X)$ entirely describes the time-to-event outcome distributions, $P(Y(a)|X, S = s)$, estimating it correctly would suffice for estimating the CATE (see (1)), even when the propensity score estimation is incorrect. We propose the following censoring-doubly-robust (CDR) signal, which enjoys this enhanced doubly-robust property:

$$\psi^{\mathsf{CDR}}_{s,a} \coloneqq \frac{\mathbf{1}\{S = s\}}{P(S = s \mid X)} \left(\frac{\mathbf{1}\{A = a\} (\psi_{s,a} - \mu_{s,a}(X))}{P(A = a \mid X, S = s)} + \mu_{s,a}(X) \right), \quad (7)$$

[1] All (in)equalities involving random variables hold almost surely throughout the manuscript.

where $\mu_{s,a}(X) := \mathsf{E}[Y|X, S=s, A=a]$ can be computed by integrating $\bar{F}_{s,a}(t|X)$ over $t \in \mathbb{R}_+$.

Theorem 1. *Suppose that Assumptions 1–4 hold and $P(S=1|X)$ is correctly estimated. Then, $\forall s, a \in \{0,1\}$*

$$\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s,a} \mid X\right] = \mathsf{E}\left[Y(a) \mid X, S=s\right],$$

if either $\bar{F}_{s,a}(t|X)$ (4), or both $\bar{G}_{s,a}(t|X)$ (3) and $P(A=a|X, S=s)$ are correctly estimated.

Although correctly estimating $\bar{F}$ under censoring remains challenging, the doubly-robust property is still desirable. For instance, consider a scenario where the censoring time tends to be high, and most of the observations are non-censored. In that case, due to the small number of censored observations, our estimates for $\bar{G}_{s,a}(t|X)$ may suffer from high variance, whereas the estimates for $\bar{F}_{s,a}(t|X)$ may be more reliable. It is also sufficient to correctly estimate $\bar{G}_{s,a}(t|X)$ and $P(A=a|X, S=s)$. This is advantageous in scenarios where further assumptions on censoring simplify the estimation of $\bar{G}$ (*e.g.*, under type-1 censoring discussed at the end of this section). We investigate the doubly-robustness of our signal empirically in Appendix C.1.

Starting from (7), we define the censoring-doubly-robust instance-wise signals for CATE in study s and "CATE difference" across studies as follows.

$$\begin{aligned} \psi^{\mathsf{CDR}}_{s} &:= \psi^{\mathsf{CDR}}_{s,a=1} - \psi^{\mathsf{CDR}}_{s,a=0}. \\ \psi^{\mathsf{CDR}} &:= \psi^{\mathsf{CDR}}_{s=1} - \psi^{\mathsf{CDR}}_{s=0}. \end{aligned} \tag{8}$$

Theorem 2. *Suppose that Assumptions 1–4 hold and $P(S=1|X)$ is correctly estimated. Then*

$$\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1} \mid X\right] = \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0} \mid X\right] = \mathrm{CATE}(X, 0),$$

where $\mathrm{CATE}(X, s)$ is defined in (1), and therefore

$$\mathsf{E}[\psi^{\mathsf{CDR}} \mid X] = 0, \tag{9}$$

if $\forall s, a \in \{0,1\}$, either $\bar{F}_{s,a}(t|X)$ (4), or both $\bar{G}_{s,a}(t|X)$ (3) and $P(A=a|X, S=s)$ are correctly estimated.

If (9) fails to hold, it means that a subset of the Assumptions 1–4 is violated. It remains to convert this condition into a hypothesis that can be tested using the maximum moment restriction via the machinery of reproducing kernel Hilbert spaces (RKHS) (Muandet et al., 2020). Before writing the full characterization of the test, we note that weaker assumptions on external validity also suffice to construct the hypothesis in (9).

Proposition 1. *Consider the same setup in Theorem 2, with the only difference being that we assume that*

$$\mathrm{CATE}(X, 0) = \mathrm{CATE}(X, 1),$$

instead of the stronger "ignorability of selection" in Assumption 2, where $\mathrm{CATE}(X, s)$ is defined in (1). Then, results in Theorem 2 continue to hold.

Proposition 1 highlights that we are effectively testing a weaker assumption. This is desirable when one is not necessarily interested in whether the conditional potential outcomes have identical distributions in both studies, but only if the CATE functions are the same. The SPRINT trial (SPRINT Research Group, 2015) is an example where only the latter holds.

Theorem 3 (MMR-based test for validity assumptions with conditionally independent censoring)**.** *Let $\psi = \psi^{\mathsf{CDR}}$ and suppose that $P(S=1|X)$ is correctly estimated. Suppose that either $\bar{F}_{s,a}(t|X)$ (4), or* both *$\bar{G}_{s,a}(t|X)$ (3) and $P(A=a|X, S=s)$ are correctly estimated $\forall s, a \in \{0,1\}$. Let $k(\cdot,\cdot)$ be an ISPD[2], continuous, and bounded kernel, and F be the RKHS endowed with $k(\cdot,\cdot)$. Suppose that $\|\mathsf{E}[\psi|X]\| < 1$ and $\mathsf{E}\left[[\psi k(X, X^0)\psi^0]^2\right] < 1$ a.s. in P_X, where (ψ, X) and (ψ^0, X^0) are i.i.d. Let $\mathsf{M} = \sup_{f \in \mathsf{F}, \|f\| \le 1} \left(\mathsf{E}[\psi f(X)]\right)^2$ be the maximum moment restriction (MMR). Then, under Assumptions 1–4, the conditional moment restriction $\mathsf{E}[\psi|X] = 0$ holds P_X a.s., which implies that the following null hypothesis H_0 holds.*

$$H_0: \ \mathsf{M}^2 = 0, \qquad H_1: \ \mathsf{M}^2 > 0.$$

We can then use the following empirical estimate of M^2 as the test statistic,

$$\hat{\mathsf{M}}^2_n = \frac{1}{n(n-1)} \sum_{i,j \in I, i \neq j} \psi_i k(x_i, x_j) \psi_j. \tag{10}$$

which has the following asymptotic distributions under the null H_0 and the alternative H_1 hypotheses.

$$\text{Under } H_0: \ \hat{\mathsf{M}}^2_n \xrightarrow{d} \sum_{j=1}^{1} \lambda_j (Z_j^2 - 1).$$

$$\text{Under } H_1: \ \sqrt{n}(\hat{\mathsf{M}}^2_n - \mathsf{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2).$$

where Z_j are i.i.d. standard normal variables and λ_j are the eigenvalues of $\psi k(x, x^0)\psi^0$, and $\sigma^2 = \mathrm{Var}_{(\psi,X)}\left(\mathsf{E}_{(\psi^0,X^0)}[\psi k(X, X^0)\psi^0]\right)$.

Theorem 3 formalizes the implications of Assumptions 1-4 as a null hypothesis we can test by calculating a statistic (see (10)) from the data and compare against a threshold t_α where $\alpha \in (0,1)$ controls the acceptable "risk" of falsifying the assumptions when they are indeed true. We provide the explicit steps in Algorithm 1.

Note that the censoring assumptions cannot be verified separately. As such, linking the rejection of the test to the violation of the validity assumptions is difficult. Nevertheless, there are cases where the censoring assumptions are true by design. Consider a study where

[2] $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be integrally strictly positive definite (ISPD) if for all $f : \mathcal{X} \to \mathbb{R}$ satisfying $0 < \|f\|_2^2 < \infty$, we have $\int_{\mathcal{X}} \int_{\mathcal{X}} f(x)k(x, x^0)f(x^0)\,\mathrm{d}x\,\mathrm{d}x^0 > 0$.

Algorithm 1 Testing for internal and external validity under conditionally independent right-censoring

Input: Combined sample from RCT $S = 0$ and OS $S = 1$: $\{X_i, A_i, S_i, \tilde{Y}_i, \Delta_i\}_{i=1}^n$, desired test level α
1. Estimate $P(S|X)$, $P(A|X, S)$, $\bar{G}$, $\bar{F}$
2. Compute ψ_i^{CDR} in (8) for $i \in \{1, \dots, n\}$
3. Compute test statistic M^2 in (10) for $\psi_i = \psi_i^{\mathsf{CDR}}$
4. Compute test threshold t_α (see Appendix C.3)
5. if $t_\alpha < \alpha$ **then** reject H_0 **else** accept H_0

patients are recruited at different times and followed until a fixed endpoint. In that case, the censoring time is known as soon as a patient enters the study (type-1 censoring (Leung et al., 1997)), and is independent from their time-to-event outcome.

3.2 Falsification with Global Censoring

While conditionally independent censoring may be plausible, it is easy to imagine settings where it is not. This makes it challenging to attribute the rejection of the test the violation of the validity assumptions rather than censoring assumptions. For instance, consider a study where patients are more likely to drop out after experiencing adverse side effects, and a short censoring time is associated with a short survival time. This induces *dependent* censoring and renders Assumption 3 implausible. Therefore, it is critical to understand how a benchmarking procedure fares under dependent censoring models (Gharari et al., 2023).

In this section, we introduce an alternative (and perhaps more plausible) censoring mechanism, which we refer to as *global censoring*, and show that the ψ^{CDR} signal in (8) can still be used to test the validity assumptions. Global censoring allows for dependent censoring, contingent on the conditional distribution of the censoring time C being identical in the RCT and the OS, reflecting the intuition that the censoring mechanism is the same in RCT and OS populations.

Assumption 5 (*Global censoring*)**.**

$$C \perp\!\!\!\perp S \mid Y, X, A.$$

CDR Signal with Global Censoring Global censoring does not entail conditionally independent censoring. Therefore, the CATE is not necessarily identifiable in the RCT or the OS. This challenges the core idea in Section 3.1 where internal validity and conditionally independent censoring imply that the instance-wise signals $\psi_{s=0}^{\mathsf{CDR}}$ and $\psi_{s=1}^{\mathsf{CDR}}$ in (8) are unbiased for the $\text{CATE}(X, 0)$ and $\text{CATE}(X, 1)$ in (1). Since $\text{CATE}(X, 0) = \text{CATE}(X, 1)$ by external validity, we proposed testing the equivalence of two signals as a proxy for testing the validity assumptions (see Theorem 3).

Nevertheless, we can show that the global censoring assumptions also imply the equivalence of $\psi_{s=0}^{\mathsf{CDR}}$ and $\psi_{s=1}^{\mathsf{CDR}}$, even if these signals are no longer unbiased for the CATE anymore. Crucially, this means that the same falsification test in Theorem 3 can also be used under global censoring, as we show next.

Theorem 4. *If Assumptions 1,2,5 hold, we have*

$$\mathsf{E}[\psi^{\mathsf{CDR}} \mid X] = \mathsf{E}[\psi_{s=1}^{\mathsf{CDR}} \quad \psi_{s=0}^{\mathsf{CDR}} \mid X] = 0, \tag{11}$$

where ψ_s^{CDR} is defined in (7,8).

Proof Sketch. Even though ψ_s^{CDR} are biased for the $\text{CATE}(X, s)$ in general (as opposed to Theorem 2), the bias in the RCT $S = 0$ and the OS $S = 1$ will be the same due to Assumption 5; therefore the conditional moment restrictions in (11) will still hold. □

Theorem 4 allows to use ψ^{CDR} to under global censoring through the same machinery in Theorem 3. This property is very desirable as it makes testing the validity assumptions, which is our original motivation, more tangible by providing a falsification test that works under two different censoring mechanisms, significantly increasing the generality of the procedure.

Alternative Signals for Global Censoring The global censoring assumption also allows the construction of more straightforward signals to test the validity assumptions. For instance, we propose the following IPW signals that use $\tilde{Y}$ in (2):

$$\psi_s^{\mathsf{IPW},\tilde{Y}} := \frac{\mathbf{1}\{S = s, A = 1\}\tilde{Y}}{P(S = s, A = 1|X)} \quad \frac{\mathbf{1}\{S = s, A = 0\}\tilde{Y}}{P(S = s, A = 0|X)}.$$
$$\psi^{\mathsf{IPW},\tilde{Y}} := \psi_{s=1}^{\mathsf{IPW},\tilde{Y}} \quad \psi_{s=0}^{\mathsf{IPW},\tilde{Y}}. \tag{12}$$

Theorem 5. *If Assumptions 1,2,5 hold, we have*

$$\mathsf{E}[\psi^{\mathsf{IPW},\tilde{Y}} \mid X] = \mathsf{E}[\psi_{s=1}^{\mathsf{IPW},\tilde{Y}} \quad \psi_{s=0}^{\mathsf{IPW},\tilde{Y}} \mid X] = 0. \tag{13}$$

$\psi_s^{\mathsf{IPW},\tilde{Y}}$ imputes censored outcomes *directly* with the censoring time. One could use another IPW signal after *dropping* the censored data, and also doubly-robust signals with additional mean outcome estimators (*e.g.*, for $\tilde{Y}$). We provide the details for alternative signals, which will also serve as "baselines" in the experiments, in Appendix D. Note that even though we call them baselines for their naive approach, these signals have not been considered for this problem before.

The weaker version of the external validity assumption *cannot* be tested under global censoring, whereas this was possible with conditionally independent censoring (see Proposition 1). Our next result formalizes this, where we consider an even stronger alternative than the exchangeability of CATEs in Proposition 1.

Proposition 2. *Consider the same setup in Theorems 4 and 5, with the only difference being we assume*

$$\mathsf{E}\left[Y(a) \mid X, S=0\right] = \mathsf{E}\left[Y(a) \mid X, S=1\right],$$

$a \in \{0, 1\}$, *instead of the stronger "ignorability of selection" in Assumption 2. Then, (11) and (13) are not true in general.*

4 IHDP EXPERIMENTS

The Infant Health and Development Program (IHDP) is an RCT that studied the effect of professional home visits on cognitive abilities in premature infants, with a sample size of 985 (Brooks-Gunn et al., 1992). We use covariate information from the IHDP trial to form semi-synthetic OS and RCT cohorts. We then simulate the binary treatment assignments, time-to-event outcomes, and censoring times based on the covariate information, as detailed in Section 4.1. The simulations cover settings with different validity assumption violations and censoring mechanisms: conditionally independent and global. Using the simulated data, we compute various signals, including the ψ^{CDR} signal in (8) and some rudimentary alternatives to serve as baselines, which are described in Section 4.2. We then conduct falsification tests for the validity assumptions using different signals as ψ signal in the 3rd step of Algorithm 1 and compare the type-1 errors and powers. Our code is available at https://github.com/demireal/censored-mmr.

4.1 Data-Generating Process

For a patient with covariates X_i in study S_i, we sample a binary treatment $A_i \sim \texttt{Bernoulli}\left(P(A=1 \mid X_i, S_i)\right)$. The propensity score $P(A \mid X, S)$ is set to a sigmoid function for both studies. Then we sample time-to-event Y_i and censoring time C_i outcomes according to the survival functions $\bar{F}_{S_i,A_i}(t \mid X_i)$ and $\bar{G}_{S_i,A_i}(t \mid X_i)$. We adopt a CoxPH framework to model the effect of covariates X and specify $\bar{F}_{S,A}(t \mid X)$ and $\bar{G}_{S,A}(t \mid X)$ as

$$\overline{W}_0(t; \lambda, p)^{\exp(X^\intercal \beta_{\text{Cox}})}, \tag{14}$$

where $\overline{W}_0(t; \lambda, p)$ is the *Weibull* baseline survival function, and the parameters λ, p, and β can be set differently for each study S and treatment group A. Exact expressions and specific parametrizations used in the experiments can be found in Appendix B.1.

4.2 Ablation Studies

We perform ablation studies to measure the efficiency of our CDR signal in (8) for testing the validity assumptions. For comparison, we propose "baselines" with no component to model censoring: IPW Y, DR Y, IPW $\tilde{Y}$, DR $\tilde{Y}$ (detailed in Appendix D). IPW Y and DR Y drop the censored data. IPW $\tilde{Y}$ and DR $\tilde{Y}$ impute the missing time-to-event with censoring time. For instance, the IPW $\tilde{Y}$ baseline uses $\psi = \psi^{\mathsf{IPW},\tilde{Y}}$ signal (12) in the 3rd step of Algorithm 1. DR baselines employ additional estimators for imputed or uncensored outcomes. We also adopt an inverse propensity of censoring-weighted (IPCW) signal that accounts for censoring by inverse-weighting with $\bar{G}$.

The significance level of the tests is set to $\alpha = 0.05$ as the desired type-1 error threshold. The synthetic RCT cohort size is the original IHDP cohort size $n_0 = 985$. We experiment with two OS cohort sizes $n_1 = 985$ and $n_1 = 2955$, where we copy the covariate data from the IHDP three times for the former. Even though the covariates are repeated, treatment and time-to-event generation processes still involve stochasticity.

4.3 Conditionally Independent Censoring

We follow the data-generating process in Section 4.1, ensuring that censoring Assumptions 3 and 4 hold and consider various violations of validity assumptions. The results are presented in Table 1 and Figure 2.

We start with setup #1, where validity assumptions hold. The baselines have very high type-1 errors, falsifying the OS despite being compatible with the RCT. Increasing the sample size does not alleviate the problem, as baselines' CATE estimates are asymptotically biased. The IPCW and CDR signals boast significantly lower type-1 errors, maintaining the test level of 0.05.

Next, we consider the violation of the external validity assumption A2; where one of the β_{Cox} parameters in (14) is different between the RCT and OS (setups #2 and #3 in Table 1, and top left in Figure 2). By $\Delta\beta_{\text{Cox}}$, we denote the magnitude of the difference, where a larger value causes a more severe violation. In addition to having high type-1 errors, IPW-based baselines also suffer from low power. This behavior can be expected, *e.g.*, if the bias introduced by naively handling the censored data cancels part of the bias from violating external validity or due to the high variance of IPW-based estimators. IPCW reacts to more severe violations of external validity; however, it cannot detect milder violations. The CDR signal enjoys higher power as a meaningful complement to its low type-1 error.

We then consider the violation of internal validity A1 in the OS by introducing "unobserved confounding" (UC) (setups #4 and #5 in Table 1 and top middle in Figure 2). We conceal the confounding covariate "sex" and adjust the violation severity through its effect on the propensity score of treatment, captured by β_{prop}. We observe that the CDR signal can detect unobserved

Table 1: *Semi-synthetic IHDP experiments with conditionally independent censoring.* Entries are the rejection rates of the null hypothesis (see Theorem 3) over 40 independent runs. $|\beta_{\text{prop}}|$ quantifies the severity of unmeasured confounding, *i.e.*, internal validity violation (A1), and $\Delta\beta_{\text{Cox}}$ the severity of external validity violation (A2).

Setup #	**1**		**2**		**3**		**4**		**5**	
Assumption validity	**A1 ✓ A2 ✓**		**A1 ✓ A2 ✗**		**A1 ✓ A2 ✗**		**A1 ✗ A2 ✓**		**A1 ✗ A2 ✓**	
Violation severity	—		$\Delta\beta_{\text{Cox}} = 0.2$		$\Delta\beta_{\text{Cox}} = 1$		$\lvert\beta_{\text{prop}}\rvert = 1$		$\lvert\beta_{\text{prop}}\rvert = 2.5$	
Metric	**Type-1 error**		**Power**		**Power**		**Power**		**Power**	
OS sample size, n_1	**985**	**2955**	**985**	**2955**	**985**	**2955**	**985**	**2955**	**985**	**2955**
DR $\tilde{Y}$	1	1	0.35	0.375	1	1	1	1	1	1
DR Y	1	1	0.55	0.6	0.85	0.95	1	1	1	1
IPW $\tilde{Y}$	1	1	0	0	0.125	0.35	1	1	1	1
IPW Y	0.9	1	0	0	0	0.025	1	1	1	1
IPCW	0	0	0	0	0.425	0.925	0.025	0.05	0.825	0.95
CDR	0	0.025	0.2	0.3	0.9	0.975	0.275	0.425	0.8	0.85

confounding, an ability pronounced by increased sample size and violation severity. As before, IPCW does not react when the violation is subtle.

Overall, falsification with CDR signal has the most reliable performance with low type-1 error and ability to detect milder violations. We also verify its doubly-robust property in Appendix C.1. Further, in Appendix C.2, we show that a "witness function" may provide explanations by revealing the regions of X where CATE estimates from RCT and OS differ the most.

4.4 Global Censoring

To simulate the global censoring mechanism, we censor patients whose time-to-event outcome exceeds a threshold by setting the censoring time to a smaller value than the threshold through the same mechanism in RCT and OS. We present the type-1 errors and powers at varying levels of validity violations in Figure 2. In contrast to the conditionally independent censoring case, all signals maintain low type-1 errors, corroborating the theory of Section 3.2. IPW and IPCW signals have lower type-1 errors than their doubly-robust counterparts; however, they are not as well-powered to detect violations of the validity assumptions.

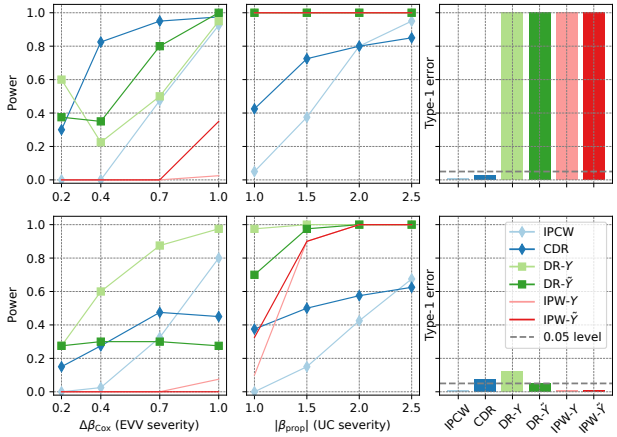

Figure 2: *Top row:* Conditionally independent censoring results. *Bottom row:* Global censoring results. EVV = External validity violation (A2). UC = Unobserved confounding (A1). OS size is $n_1 = 2955$.

5 WHI EXPERIMENTS

The Women's Health Initiative (WHI) was launched in the early 1990s to study various health outcomes in postmenopausal women. Previous studies noted discrepancies between the RCT and OS components of the WHI (Prentice et al., 2005). We focus on the effect of combination hormone therapy on a composite outcome: the minimum time-to-event among multiple endpoints such as heart failure, cancer, and death.

Similar to Hussain et al. (2023), we limit the follow-up to seven years since the treatment assignment. While previous studies discarded censoring by binarizing the time-to-event outcome and imputing the outcome for censored patients with $Y = 0$, our framework allows us to use the true outcomes and explicitly model the censoring component. We used 954 features available in the RCT and OS and performed principal component (PC) analysis to alleviate collinearity-related instabilities in Cox regression models (350 PCs, capturing 90% of the variance, details in Appendix B.2). We split the data into ten folds, estimate the nuisance functions using nine folds (step 1 in Algorithm 1), and perform the MMR test with the remaining fold. We repeat ten times and report the average rejection rates.

Table 2: WHI experiments. Average rejection rate over 10-folds for different fractions of selection biases f.

	IPW $\tilde{Y}$	DR $\tilde{Y}$	IPCW	CDR
$f = 0$	0	0.5	0	0.6
$f = 0.1$	0.1	0.2	0.1	0.9
$f = 0.25$	0	0.4	0.3	0.9

We also simulate selection bias by dropping $f \in (0, 1)$ fraction of patients from the RCT's control group that experienced an event. The results are presented in Table 2. IPW-based methods have lower rejection rates, while the CDR signal has the highest rejection rate, which increases with the selection bias f.

6 RELATED WORK IN EPIDEMIOLOGY

Evaluating the internal and external validity of RCTs and OSes has been of significant interest in the epidemiology literature. Here, we surface some references for the interested reader.

One elegant concept is *negative outcomes* (Lipsitch et al., 2010; Sofer et al., 2016). A negative outcome is known or expected to be unaffected by the treatment. Therefore, a significant difference in a negative outcome between the treatment and control groups may point to flaws in the study design. For instance, Dagan et al. (2021) leverage the fact that the COVID-19 vaccine should not have a significant effect within the first few days of administration to guide their adjustment for confounders in the observational data.

Viele et al. (2014) investigate methods to incorporate *historical controls* into the design of new RCTs to make them more efficient. Their "test-then-pool" procedure first compares the historical controls to trial controls before pooling them. Hartman et al. (2015) study estimating the *population* average treatment effect *on the treated* from an RCT, where the population of interest is defined by an OS cohort. Their approach involves a placebo test in the first step to gauge the generalizability of the trial to the target population. de Luna and Johansson (2014) show how an alternative set of causal assumptions on *instrumental variables* can be used to construct a test for the no unmeasured confounding assumption in an OS.

We close by noting Forbes and Dahabreh (2020) and Wang et al. (2023). They provide thorough empirical evidence regarding the compatibility of RCTs and their OS counterparts by analyzing complementary findings from an extensive set of studies in the literature.

7 CONCLUSION

We developed a framework to test the validity of an OS by benchmarking it against an RCT, when the outcomes are right-censored. We considered the common conditionally independent censoring condition and introduced a novel one: global censoring. We demonstrated that naively handling the censoring leads to unreliable tests. In contrast, our censoring-doubly-robust signal facilitated tests with low type-1 error and high sensitivity to violations of the validity assumptions under both censoring scenarios, making it a promising candidate for generally applicable benchmarking procedures under different censoring scenarios.

Acknowledgments

The authors thank the anonymous reviewers for their helpful suggestions and Ming-Chieh Shih for discussions during the earlier versions of the manuscript. ID was supported by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. ZH was supported by an ASPIRE award from The Mark Foundation for Cancer Research and by the National Cancer Institute of the National Institutes of Health under Award Number F30CA268631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MO and DS were supported in part by Office of Naval Research Award No. N00014-21- 1-2807. EDB was funded by a FWO-SB PhD grant. This manuscript was prepared using WHI-CTOS Research Materials obtained from the National Heart, Lung, and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the WHI-CTOS or the NHLBI.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3): 350–359, 1992.

Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 133–145, 2021.

Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Com-*

puter Methods and Programs in Biomedicine, 75(1): 45–49, 2004.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34:26740–26753, 2021.

Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A Katz, Miguel A Hernán, Marc Lipsitch, Ben Reis, and Ran D Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 2021.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75 (2):685–694, June 2019.

Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39 (14):1999–2014, 2020a.

Issa J Dahabreh, James M Robins, and Miguel A Hernán. Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology*, 31(5):614–619, 2020b.

Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019. URL https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html.

Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *arXiv preprint arXiv:2312.03871*, 2023.

Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2(2):187–199, 2014.

Shaun P Forbes and Issa J Dahabreh. Benchmarking observational analyses against randomized trials: a review of studies assessing propensity score methods. *Journal of general internal medicine*, 35:1396–1404, 2020.

Boris Gershman, David P Guo, and Issa J Dahabreh. Using observational data for personalized medicine when clinical trial evidence is limited. *Fertility and Sterility*, 109(6):946–951, 2018.

Ali Hossein Foomani Gharari, Michael Cooper, Russell Greiner, and Rahul G Krishnan. Copula-based deep survival models for dependent censoring. In *Uncertainty in Artificial Intelligence*, pages 669–680. PMLR, 2023.

Government of Canada. Optimizing the use of real world evidence to inform regulatory decision-making, 2019. URL https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/announcements/optimizing-real-world-evidence-regulatory-decisions.html.

Harrison J Hansford, Aidan G Cashin, Matthew D Jones, Sonja A Swanson, Nazrul Islam, Susan RG Douglas, Rodrigo RN Rizzo, Jack J Devonshire, Sam A Williams, Issa J Dahabreh, et al. Reporting of observational studies explicitly aiming to emulate randomized trials: A systematic review. *JAMA Network Open*, 6(9):e2336023–e2336023, 2023.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A,*, 178(3):757–778, June 2015.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

Miguel A Hernan and James M Robins. *Causal Inference*. CRC Press, Boca Raton, FL, February 2021.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

Miguel A Hernán, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Meir J Stampfer, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6): 766, 2008.

Miguel A Hernán, James M Robins, et al. Per-protocol analyses of pragmatic trials. *N Engl J Med*, 377(14): 1391–1398, 2017.

Zeshan Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. *arxiv preprint arXiv:2209.13708*, 2022.

Zeshan Hussain, Ming-Chieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *International Conference on Artificial Intelligence and Statistics*, pages 5869–5898, 2023.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

Rickard Karlsson and Jesse Krijthe. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, 36, 2024.

Kwan-Moon Leung, Robert M Elashoff, and Abdelmonem A Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1):83–104, 1997.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

Sara Lodi, Andrew Phillips, Jens Lundgren, Roger Logan, Shweta Sharma, Stephen R Cole, Abdel Babiker, Matthew Law, Haitao Chu, Dana Byrne, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. *American Journal of Epidemiology*, 188(8):1569–1577, 2019.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.

NICE. Nice real-world evidence framework, 2022. URL https://www.nice.org.uk/corporate/ecd9/chapter/overview.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

Ross L Prentice, Robert Langer, Marcia L Stefanick, Barbara V Howard, Mary Pettinger, Garnet Anderson, David Barad, J David Curb, Jane Kotchen, Lewis Kuller, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *American journal of epidemiology*, 162(5):404–414, 2005.

Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

Daniel Rubin and Mark J van der Laan. A doubly robust censoring unbiased transformation. *The international journal of biostatistics*, 3(1), 2007.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Tamar Sofer, David B Richardson, Elena Colicino, Joel Schwartz, and Eric J Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

Therapeutic Goods Administration. Real world evidence and patient reported outcomes in the regulatory context, 2023. URL https://www.tga.gov.au/real-world-evidence-rwe-and-patient-reported-outcomes-pros.

Ingrid Van Keilegom, Michael G Akritas, and Noël Veraverbeke. Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics & Data Analysis*, 35 (4):487–500, 2001.

Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

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

Shirley V Wang, Sebastian Schneeweiss, Jessica M Franklin, Rishi J Desai, William Feldman, Elizabeth M Garry, Robert J Glynn, Kueiyu Joshua Lin, Julie Paik, Elisabetta Patorno, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. *JAMA*, 329(16):1376–1385, 2023.

Kabir Yadav and Roger J Lewis. Immortal time bias in observational studies. *Jama*, 325(7):686–687, 2021.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. See Sections 2 and 3.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. See Section 3.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes. Link to the code repository is provided in Section 4.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes. All the assumptions and theoretical results are listed in the main paper in Sections 2 and 3.]
   (b) Complete proofs of all theoretical results. [Yes. All the proofs are included in Appendix A.]
   (c) Clear explanations of any assumptions. [Yes. The motivations/insights behind the assumptions are listed either before or after stating the critical assumptions.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. See the repository link in Section 4]
   (b) All the training details (*e.g.*, data splits, hyperparameters, how they were chosen). [Yes. See Sections 4, 5 and Appendix B]
   (c) A clear definition of the specific measure or statistics and error bars (*e.g.*, with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (*e.g.*, type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (*e.g.*, code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [See the repository link in Section 4]
   (d) Information about consent from data providers/curators. [Yes. See Appendix B.2]
   (e) Discussion of sensible content if applicable, *e.g.*, personally identifiable information or offensive content. [Yes. See Appendix B.2]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Table of Contents

A PROOFS

Lemma A.1. *Suppose that Assumption 1 holds. We have,* $\forall s, a \in \{0,1\}$

$$P(Y(a)|X, S = s) = P(Y|X, S = s, A = a)$$

Proof.

$$\begin{aligned} P(Y(a)|X, S = s) &= P(Y(a)|X, A = a, S = s) \\ &= P(Y|X, A = a, S = s) \end{aligned}$$

by no unobserved confounding and consistency. □

A.1 Lemma 1

Lemma 1. *Suppose that Assumptions 1,3, and 4 hold. From Theorem 1 in Rubin and van der Laan (2007) and Assumption 1, we have,* $\forall s, a \in \{0,1\}$,

$$\mathbb{E}\left[\psi_{s,a} \mid X, S = s, A = a\right] = \mathbb{E}\left[Y(a) \mid X, S = s\right],$$

if $\bar{G}_{s,a}(t|X)$ (3) *or* $\bar{F}_{s,a}(t|X)$ (4) *is correctly estimated.*

Proof.

$$\begin{aligned}\mathsf{E}\left[\psi_{s,a} \mid X, S=s, A=a\right] &= \mathsf{E}\left[Y \mid X, S=s, A=a\right] && \text{(Theorem 1 in Rubin and van der Laan (2007))}\\ &= \mathsf{E}\left[Y(a) \mid X, S=s\right] && \text{(Lemma A.1)}\end{aligned}$$

□

A.2 Corollary 1

Corollary 1. *Suppose that Assumptions 1,3,4 hold and $P(S=1 \mid X)$ is correctly estimated. Let*

$$\psi_{s,a}^{\mathsf{IPW},} := \frac{\mathbf{1}\{S=s, A=a\}\psi_{s,a}}{P(S=s \mid X)P(A=a \mid X, S=s)}. \tag{6}$$

Then $\forall s, a \in \{0,1\}$, $\mathsf{E}\left[\psi_{s,a}^{\mathsf{IPW},} \mid X\right] = \mathsf{E}\left[Y(a) \mid X, S=s\right]$ if $P(A=a \mid X, S=s)$, and either $\bar{G}_{s,a}(t \mid X)$ (3) or $\bar{F}_{s,a}(t \mid X)$ (4) are correctly estimated.

Proof. Let us denote by $\hat{P}(A=a \mid X, S=s)$ the estimated propensity score.

$$\begin{aligned}&\mathsf{E}\left[\frac{\mathbf{1}\{S=s, A=a\}\psi_{s,a}}{P(S=s \mid X)\hat{P}(A=a \mid X, S=s)} \,\middle|\, X\right]\\ &= \mathsf{E}\left[\frac{\psi_{s,a}}{P(S=s \mid X)\hat{P}(A=a \mid X, S=s)} \,\middle|\, X, S=s, A=a\right] P(S=s, A=a \mid X)\\ &= \mathsf{E}\left[\psi_{s,a} \mid X, S=s, A=a\right]\\ &= \mathsf{E}\left[Y(a) \mid X, S=s\right] && \text{(Lemma 1)}\end{aligned}$$

where the probability terms cancel out since $\hat{P}(A=a \mid X, S=s) = P(A=a \mid X, S=s)$ is correctly estimated. □

A.3 Theorem 1

Theorem 1. *Suppose that Assumptions 1–4 hold and $P(S=1 \mid X)$ is correctly estimated. Then, $\forall s, a \in \{0,1\}$*

$$\mathsf{E}\left[\psi_{s,a}^{\mathsf{CDR}} \mid X\right] = \mathsf{E}\left[Y(a) \mid X, S=s\right],$$

if either $\bar{F}_{s,a}(t \mid X)$ (4), or both $\bar{G}_{s,a}(t \mid X)$ (3) and $P(A=a \mid X, S=s)$ are correctly estimated.

Proof. Let us denote the estimates for the nuisance functions with

$$\hat{P}(A=a \mid X, S=s), \quad \hat{G}_{s,a}(t \mid X), \quad \hat{F}_{s,a}(t \mid X), \quad \hat{\mu}_{s,a}(X)$$

and we have

$$\hat{\psi}_{s,a}^{\mathsf{CDR}} = \frac{\mathbf{1}\{S=s\}}{P(S=s \mid X)} \left(\frac{\mathbf{1}\{A=a\}(\hat{\psi}_{s,a} - \hat{\mu}_{s,a}(X))}{\hat{P}(A=a \mid X, S=s)} + \hat{\mu}_{s,a}(X)\right)$$

First, assume that the survival function of the time-to-event outcome is correctly estimated but the survival function of the censoring time and the propensity score are not. That is,

$$\hat{F}_{s,a}(t \mid X) = \bar{F}_{s,a}(t \mid X) \tag{15}$$
$$\hat{\mu}_{s,a}(X) = \mu_{s,a}(X) \tag{16}$$
$$\hat{G}_{s,a}(t \mid X) \neq \bar{G}_{s,a}(t \mid X)$$
$$\hat{P}(A=a \mid X, S=s) \neq P(A=a \mid X, S=s)$$

Note that

$$\mu_{s,a}(X) = \mathsf{E}\left[Y \mid X, S=s, A=a\right]$$

$$= \mathsf{E}\left[Y(a) \mid X, S = s\right]$$
$$= \mathsf{E}\left[\hat{\psi}_{s,a} \mid X, S = s, A = a\right] \tag{17}$$

where the second equality is by Lemma A.1 and the third by Lemma 1 thanks to (15). We have

$$\begin{aligned}
\mathsf{E}\left[\hat{\psi}^{\mathsf{CDR}}_{s,a} \mid X\right] &= \mathsf{E}\left[\frac{\mathbf{1}\{S = s, A = a\}\left(\hat{\psi}_{s,a} \quad \hat{\mu}_{s,a}(X)\right)}{P(S = s \mid X)\hat{P}(A = a \mid X, S = s)} \middle| X\right] + \mathsf{E}\left[\frac{\mathbf{1}\{S = s\}\hat{\mu}_{s,a}(X)}{P(S = s \mid X)} \middle| X\right] \\
&= \mathsf{E}\left[\hat{\psi}_{s,a} \quad \hat{\mu}_{s,a}(X) \mid X, S = s, A = a\right]\frac{P(A = a \mid X, S = s)}{\hat{P}(A = a \mid X, S = s)} + \hat{\mu}_{s,a}(X) \\
&= \underbrace{\left(\mathsf{E}\left[\hat{\psi}_{s,a} \mid X, S = s, A = a\right] \quad \mu_{s,a}(X)\right)}_{0 \text{ by } (17)}\frac{P(A = a \mid X, S = s)}{\hat{P}(A = a \mid X, S = s)} + \mu_{s,a}(X) \qquad (18) \\
&= \mathsf{E}\left[Y(a) \mid X, S = s\right] \qquad \text{(Lemma A.1)}
\end{aligned}$$

where (18) follows from (16). Next, assume that the propensity score and the survival function of the censoring time is correctly estimated, but the survival function of the outcome is not,

$$\hat{P}(A = a \mid X, S = s) = P(A = a \mid X, S = s) \tag{19}$$
$$\hat{G}(t \mid X, S = s, A = a) = \bar{G}(t \mid X, S = s, A = a) \tag{20}$$
$$\hat{F}(t \mid X, S = s, A = a) \neq \bar{F}(t \mid X, S = s, A = a)$$

We have

$$\begin{aligned}
\mathsf{E}\left[\hat{\psi}^{\mathsf{CDR}}_{s,a} \mid X\right] &= \mathsf{E}\left[\frac{\mathbf{1}\{S = s, A = a\}\hat{\psi}_{s,a}}{P(S = s \mid X)\hat{P}(A = a \mid X, S = s)} \quad \frac{\mathbf{1}\{S = s, A = a\}\hat{\mu}_{s,a}(X)}{P(S = s \mid X)\hat{P}(A = a \mid X, S = s)} + \frac{\mathbf{1}\{S = s\}\hat{\mu}_{s,a}(X)}{P(S = s \mid X)} \middle| X\right] \\
&= \mathsf{E}\left[\frac{\mathbf{1}\{S = s, A = a\}\hat{\psi}_{s,a}}{P(S = s, A = a \mid X)} \middle| X\right] \quad \mathsf{E}\left[\frac{\mathbf{1}\{S = s\}\hat{\mu}_{s,a}(X)}{P(S = s \mid X)}\left(\frac{\mathbf{1}\{A = a\} \quad P(A = a \mid X, S = s)}{P(A = a \mid X, S = s)}\right) \middle| X\right] \qquad (21) \\
&= \mathsf{E}\left[\hat{\psi}_{s,a} \mid X, S = s, A = a\right] \quad \hat{\mu}_{s,a}(X) \quad \underbrace{\mathsf{E}\left[\left(\frac{\mathbf{1}\{A = a\} \quad P(A = a \mid X, S = s)}{P(A = a \mid X, S = s)}\right) \middle| X, S = s\right]}_{0} \\
&= \mathsf{E}\left[Y(a) \mid X, S = s\right] \qquad (22)
\end{aligned}$$

where we have (21) by (19), and (22) follows from Lemma 1 since we have (20). □

A.4 Theorem 2

Theorem 2. *Suppose that Assumptions 1–4 hold and $P(S = 1 \mid X)$ is correctly estimated. Then*

$$\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1} \mid X\right] = \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0} \mid X\right] = \mathrm{CATE}(X, 0),$$

where $\mathrm{CATE}(X, s)$ *is defined in* (1)*, and therefore*

$$\mathsf{E}[\psi^{\mathsf{CDR}} \mid X] = 0, \tag{9}$$

if $\forall s, a \in \{0, 1\}$*, either* $\bar{F}_{s,a}(t \mid X)$ (4)*, or both* $\bar{G}_{s,a}(t \mid X)$ (3) *and* $P(A = a \mid X, S = s)$ *are correctly estimated.*

Proof. Note that $\mathrm{CATE}(X, 0) = \mathsf{E}\left[Y(1) \quad Y(0) \mid X, S = 0\right]$.

$$\begin{aligned}
\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1} \mid X\right] &= \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a=1} \mid X\right] \quad \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a=0} \mid X\right] \\
&= \mathsf{E}\left[Y(1) \mid X, S = 1\right] \quad \mathsf{E}\left[Y(0) \mid X, S = 1\right] && \text{(Theorem 1)} \\
&= \mathsf{E}\left[Y(1) \mid X, S = 0\right] \quad \mathsf{E}\left[Y(0) \mid X, S = 0\right] && \text{(Ignorability of selection, Assumption 2)} \\
&= \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0} \mid X\right] && \text{(by symmetry)}
\end{aligned}$$

□

A.5 Proposition 1

Proposition 1. *Consider the same setup in Theorem 2, with the only difference being that we assume that*

$$\mathrm{CATE}(X,0) = \mathrm{CATE}(X,1),$$

instead of the stronger "ignorability of selection" in Assumption 2, where $\mathrm{CATE}(X,s)$ *is defined in (1). Then, results in Theorem 2 continue to hold.*

Proof. The proof is identical to the proof of Theorem 2, only difference being we invoke the alternative external validity assumption $\mathrm{CATE}(X,0) = \mathrm{CATE}(X,1)$ in the third step, instead of the stronger ignorability of selection. □

A.6 Theorem 3

Theorem 3 (MMR-based test for validity assumptions with conditionally independent censoring)**.** *Let* $\psi = \psi^{\mathsf{CDR}}$ *and suppose that* $P(S=1|X)$ *is correctly estimated. Suppose that either* $\bar{F}_{s,a}(t|X)$ *(4), or* both $\bar{G}_{s,a}(t|X)$ *(3) and* $P(A=a|X,S=s)$ *are correctly estimated* $\forall s,a \in \{0,1\}$*. Let* $k(\cdot,\cdot)$ *be an ISPD*[3]*, continuous, and bounded kernel, and* $\mathcal{F}$ *be the RKHS endowed with* $k(\cdot,\cdot)$*. Suppose that* $|\mathbb{E}[\psi|X]| < 1$ *and* $\mathbb{E}\left[[\psi k(X,X^0)\psi^0]^2\right] < 1$ *a.s. in* P_X*, where* (ψ, X) *and* (ψ^0, X^0) *are i.i.d. Let* $\mathsf{M} = \sup_{f\in\mathcal{F},\|f\|\leq 1}\left(\mathbb{E}\left[\psi f(X)\right]\right)^2$ *be the maximum moment restriction (MMR). Then, under Assumptions 1–4, the conditional moment restriction* $\mathbb{E}[\psi|X] = 0$ *holds* P_X *a.s., which implies that the following null hypothesis* H_0 *holds.*

$$H_0: \ \mathsf{M}^2 = 0, \qquad H_1: \ \mathsf{M}^2 > 0.$$

We can then use the following empirical estimate of M^2 *as the test statistic,*

$$\hat{\mathsf{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j\in I, i\neq j} \psi_i k(x_i,x_j)\psi_j. \tag{10}$$

which has the following asymptotic distributions under the null H_0 *and the alternative* H_1 *hypotheses.*

$$\text{Under } H_0: \ \hat{\mathsf{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1).$$

$$\text{Under } H_1: \ \sqrt{n}(\hat{\mathsf{M}}_n^2 - \mathsf{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2).$$

where Z_j *are i.i.d. standard normal variables and* λ_j *are the eigenvalues of* $\psi k(x,x^0)\psi^0$*, and* $\sigma^2 = \mathrm{Var}_{(\psi,X)}\left(\mathbb{E}_{(\psi^0,X^0)}[\psi k(X,X^0)\psi^0]\right)$.

Proof. By Theorem 2 and (8), we have

$$\mathbb{E}[\psi|X] = \mathbb{E}\left[\psi^{\mathsf{CDR}}|X\right] = 0$$

The hypothesis test results then follow from Theorem 3.1 in Hussain et al. (2023). □

Lemma A.2. *Suppose that Assumptions 1, 2, 5 hold. We have, for all* $a \in \{0,1\}$ *and* $t \in \mathbb{R}_+$

$$\bar{G}_{s=0,a}(t|X) = \bar{G}_{s=1,a}(t|X)$$
$$Q_{s=0,a}(X,t) = Q_{s=1,a}(X,t)$$

Proof. Let us start with the first statement.

$$\bar{G}_{s=0,a}(t|X) = P(C > t|X, S=0, A=a)$$

[3] $k(\cdot,\cdot): \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be integrally strictly positive definite (ISPD) if for all $f: \mathcal{X} \to \mathbb{R}$ satisfying $0 < \|f\|_2^2 < \infty$, we have $\int_{\mathcal{X}}\int_{\mathcal{X}} f(x)k(x,x^0)f(x^0)\,\mathrm{d}x\,\mathrm{d}x^0 > 0$.

$$
\begin{aligned}
&= \sum_y P(C > t | Y = y, X, S = 0, A = a) P(Y = y | X, S = 0, A = a) \\
&= \sum_y P(C > t | Y = y, X, S = 0, A = a) P(Y(a) = y | X, S = 0) && \text{(Lemma A.1)} \\
&= \sum_y P(C > t | Y = y, X, S = 1, A = a) P(Y(a) = y | X, S = 1) && \text{(Assumptions 2, 5)} \\
&= \bar{G}_{s=1,a}(t | X) && \text{(by symmetry)}
\end{aligned}
$$

For the second statement, we write

$$
\begin{aligned}
Q_{s=0,a}(X, t) &= \mathsf{E}\left[Y | X, Y > t, S = 0, A = a\right] \\
&= \sum_y y P(Y = y | X, Y > t, S = 0, A = a) \\
&= \sum_y y \frac{P(Y(a) = y, Y(a) > t | X, S = 0)}{P(Y(a) > t | X, S = 0)} && \text{(Lemma A.1)} \\
&= \sum_y y \frac{P(Y(a) = y, Y(a) > t | X, S = 1)}{P(Y(a) > t | X, S = 1)} && \text{(Assumption 2)} \\
&= Q_{s=1,a}(X, t) && \text{(by symmetry)}
\end{aligned}
$$

□

Lemma A.3. *Suppose that Assumptions 1, 2, 5 hold. We have*

$$
P(Y \quad C | X, S = 0, A) = P(Y \quad C | X, S = 1, A)
$$

Proof.

$$
\begin{aligned}
P(Y \quad C | X, S = 0, A) &= \sum_y P(y \quad C | Y = y, X, S = 0, A) P(Y = y | X, S = 0, A) \\
&= \sum_y P(y \quad C | Y = y, X, S = 0, A) P(Y(A) = y | X, S = 0) && \text{(Lemma A.1)} \\
&= \sum_y P(y \quad C | Y = y, X, S = 1, A) P(Y(A) = y | X, S = 1) && \text{(Assumptions 2, 5)} \\
&= P(Y \quad C | X, S = 1, A) && \text{(by symmetry)}
\end{aligned}
$$

□

Lemma A.4. *Suppose that Assumptions 1, 2, 5 hold. We have*

$$
\begin{aligned}
\mathsf{E}\left[Y | X, S = 0, A, Y \quad C\right] &= \mathsf{E}\left[Y | X, S = 1, A, Y \quad C\right] \\
\mathsf{E}\left[C | X, S = 0, A, Y > C\right] &= \mathsf{E}\left[C | X, S = 1, A, Y > C\right]
\end{aligned}
$$

Proof. For the first statement we write

$$
\begin{aligned}
\mathsf{E}\left[Y | X, S = 0, A, Y \quad C\right] &= \sum_y y P(Y = y | X, S = 0, A, Y \quad C) \\
&= \sum_y y \frac{P(Y \quad C | Y = y, X, S = 0, A) P(Y = y | X, S = 0, A = a)}{P(Y \quad C | X, S = 0, A)} \\
&= \sum_y y \frac{P(y \quad C | Y = y, X, S = 0, A) P(Y(A) = y | X, S = 0)}{P(Y \quad C | X, S = 0, A)} && \text{(Lemma A.1)} \\
&= \sum_y y \frac{P(y \quad C | Y = y, X, S = 1, A) P(Y(A) = y | X, S = 1)}{P(Y \quad C | X, S = 1, A)} && \text{(Lemma A.3, Assumptions 2,5)}
\end{aligned}
$$

$$= \mathsf{E}\left[Y \middle| X, S=1, A, Y \quad C\right] \qquad \text{(by symmetry)}$$

For the second statement we write

$$\begin{aligned}
&\mathsf{E}\left[C \middle| X, S=0, A, Y > C\right] \\
=&\sum_y \mathsf{E}\left[C \middle| X, S=0, A, Y > C, Y = y\right] P(Y = y \middle| X, S=0, A, Y > C) \\
=&\sum_y \mathsf{E}\left[C \middle| X, S=0, A, Y > C, Y = y\right] \frac{P(y > C \middle| Y = y, X, S=0, A) P(Y = y \middle| X, S=0, A)}{P(Y > C \middle| X, S=0, A)} \\
=&\sum_y \mathsf{E}\left[C \middle| X, S=0, A, Y > C, Y = y\right] \frac{P(y > C \middle| Y = y, X, S=0, A) P(Y(A) = y \middle| X, S=0)}{P(Y > C \middle| X, S=0, A)} \qquad \text{(Lemma A.1)} \\
=&\sum_y \mathsf{E}\left[C \middle| X, S=0, A, Y > C, Y = y\right] \underbrace{\frac{P(y > C \middle| Y = y, X, S=1, A) P(Y(A) = y \middle| X, S=1)}{P(Y > C \middle| X, S=1, A)}}_{=:g(X,A,Y=y,S=1)} \\
& \qquad \text{(Lemma A.3, Assumptions 2, 5)} \\
=&\sum_y g(X, A, Y = y, S=1) \sum_{c: y > c} c P(C = c \middle| X, S=0, A, Y > C, Y = y) \\
=&\sum_y g(X, A, Y = y, S=1) \sum_{c: y > c} c \frac{P(y > C, C = c \middle| X, S=0, A, Y = y)}{P(y > C \middle| X, S=0, A, Y = y)} \\
=&\sum_y g(X, A, Y = y, S=1) \sum_{c: y > c} c \frac{\overbrace{P(y > c \middle| X, S=0, A, C = c, Y = y)}^{=1} P(C = c \middle| X, S=0, A, Y = y)}{P(y > C \middle| X, S=0, A, Y = y)} \\
=&\sum_{y, c: y > c} g(X, A, Y = y, S=1) \frac{c P(C = c \middle| X, S=1, A, Y = y)}{P(y > C \middle| X, S=1, A, Y = y)} \qquad \text{(Assumption 5)} \\
=&\mathsf{E}\left[C \middle| X, S=1, A, Y > C\right] \qquad \text{(by symmetry)}
\end{aligned}$$

□

A.7 Theorem 4

Theorem 4. *If Assumptions 1,2,5 hold, we have*

$$\mathsf{E}[\psi^{\mathsf{CDR}} \mid X] = \mathsf{E}[\psi^{\mathsf{CDR}}_{s=1} \quad \psi^{\mathsf{CDR}}_{s=0} \mid X] = 0, \tag{11}$$

where ψ^{CDR}_s is defined in (7,8).

Proof. If we can show that $\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0,a} \middle| X\right] = \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a} \middle| X\right]$ for all $a \in \{0,1\}$, we are done since $\mathsf{E}\left[\psi^{\mathsf{CDR}}_s \middle| X\right] = \mathsf{E}\left[\psi^{\mathsf{CDR}}_{s,a=1} \quad \psi^{\mathsf{CDR}}_{s,a=0} \middle| X\right]$ for all $s \in \{0,1\}$.

Recall that $\psi^{\mathsf{CDR}}_{s,a}$ is defined in Eq. (7) as

$$\psi^{\mathsf{CDR}}_{s,a} := \frac{\mathbf{1}\{S = s\}}{P(S = s \mid X)} \quad \left(\frac{\mathbf{1}\{A = a\}(\psi_{s,a} \quad \mu_{s,a}(X))}{P(A = a \mid X, S = s)} + \mu_{s,a}(X)\right)$$

Allowing us to write

$$\begin{aligned}
&\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0,a} \middle| X\right] \\
=&\mathsf{E}\left[\frac{\mathbf{1}\{S = 0, A = a\}\psi_{s=0,a}}{P(S = 0, A = a \middle| X)} \middle| X\right] \quad \underbrace{\cancel{\mathsf{E}\left[\frac{\mathbf{1}\{S = 0, A = a\}\mu_{s=0,a}(X)}{P(S = 0, A = a \middle| X)} \middle| X\right]}}_{\mu_{s=0,a}(X)} + \underbrace{\cancel{\mathsf{E}\left[\frac{\mathbf{1}\{S = 0\}\mu_{s=0,a}(X)}{P(S = 0 \middle| X)} \middle| X\right]}}_{\mu_{s=0,a}(X)} \\
=&\mathsf{E}\left[\psi_{s=0,a} \middle| X, S = 0, A = a\right]
\end{aligned}$$

$$= \mathbb{E}\left[\underbrace{\frac{\mathbf{1}\, f\Delta = 1g\, Y}{\bar{G}_{s=0,a}(YjX)}}_{\text{Term 1}} + \underbrace{\frac{\mathbf{1}\, f\Delta = 0g\, Q_{s=0,a}(X,C)}{\bar{G}_{s=0,a}(CjX)}}_{\text{Term 2}} \quad \underbrace{\int_{1}^{\tilde{Y}} \frac{Q_{s=0,a}(X,c)}{\bar{G}^2_{s=0,a}(cjX)} dG_{s=0,a}(cjX)}_{\text{Term 3}} \Big| X, S=0, A=a \right] \tag{23}$$

where in the last line, we have substituted the definition of $\psi_{s=0,a}$ from Eq. (5). We will proceed separately for each term. Note that $\mathbf{1}\, f\Delta = 1g = \mathbf{1}\, fY \quad Cg$.

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}\, f\Delta = 1g\, Y}{\bar{G}_{s=0,a}(YjX)} \Big| X, S=0, A=a \right] \\
=&\mathbb{E}\left[\frac{Y}{\bar{G}_{s=0,a}(YjX)} \Big| X, S=0, A=a, Y \quad C \right] P(Y \quad CjX, S=0, A=a) \\
=&\sum_y \frac{y}{\bar{G}_{s=0,a}(yjX)} P(Y=yjX, S=0, A=a, Y \quad C) P(Y \quad CjX, S=0, A=a) \\
=&\sum_y \frac{y}{\bar{G}_{s=0,a}(yjX)} P(Y \quad CjY=y, X, S=0, A=a) P(Y=yjX, S=0, A=a) \\
=&\sum_y \frac{y}{\bar{G}_{s=0,a}(yjX)} P(y \quad CjY=y, X, S=0, A=a) P(Y(a)=yjX, S=0) && \text{(Lemma A.1)} \\
=&\sum_y \frac{y}{\bar{G}_{s=1,a}(yjX)} P(y \quad CjY=y, X, S=1, A=a) P(Y(a)=yjX, S=1) && \text{(Lemma A.2, Assumptions 2, 5)} \\
=&\mathbb{E}\left[\frac{\mathbf{1}\, f\Delta = 1g\, Y}{\bar{G}_{s=1,a}(YjX)} \Big| X, S=1, A=a \right] && \text{(by symmetry)}
\end{aligned}$$

We continue with Term 2

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}\, f\Delta = 0g\, Q_{s=0,a}(X,C)}{\bar{G}_{s=0,a}(CjX)} \Big| X, S=0, A=a \right] \\
=&\mathbb{E}\left[\frac{Q_{s=0,a}(X,C)}{\bar{G}_{s=0,a}(CjX)} \Big| X, S=0, A=a, Y>C \right] P(Y>CjX, S=0, A=a) \\
=&\sum_c \frac{Q_{s=0,a}(X,c)}{\bar{G}_{s=0,a}(cjX)} P(C=cjX, S=0, A=a, Y>C) P(Y>CjX, S=0, A=a) \\
=&\sum_c \frac{Q_{s=1,a}(X,c)}{\bar{G}_{s=1,a}(cjX)} P(C=c, Y>CjX, S=0, A=a) && \text{(Lemma A.2)} \\
=&\sum_c \frac{Q_{s=1,a}(X,c)}{\bar{G}_{s=1,a}(cjX)} \sum_y P(C=c, Y>CjY=y, X, S=0, A=a) P(Y=yjX, S=0, A=a) \\
=&\sum_c \frac{Q_{s=1,a}(X,c)}{\bar{G}_{s=1,a}(cjX)} \sum_y P(C=c, y>CjY=y, X, S=0, A=a) P(Y(a)=yjX, S=0) && \text{(Lemma A.1)} \\
=&\sum_c \frac{Q_{s=1,a}(X,c)}{\bar{G}_{s=1,a}(cjX)} \sum_y P(C=c, y>CjY=y, X, S=1, A=a) P(Y(a)=yjX, S=1) && \text{(Assumptions 2, 5)} \\
=&\mathbb{E}\left[\frac{\mathbf{1}\, f\Delta = 0g\, Q_{s=1,a}(X,C)}{\bar{G}_{s=1,a}(CjX)} \Big| X, S=1, A=a \right] && \text{(by symmetry)}
\end{aligned}$$

We continue with Term 3

$$\begin{aligned}
&\mathbb{E}\left[\int_{1}^{\tilde{Y}} \frac{Q_{s=0,a}(X,c)}{\bar{G}^2_{s=0,a}(cjX)} dG_{s=0,a}(cjX) \Big| X, S=0, A=a \right] \\
=&\sum_{y,c} \mathbb{E}\left[\int_{1}^{\tilde{y}} \frac{Q_{s=0,a}(X,c)}{\bar{G}^2_{s=0,a}(cjX)} dG_{s=0,a}(cjX) \Big| Y=y, C=c, X, S=0, A=a \right] P(Y=y, C=cjX, S=0, A=a) \\
=&\sum_{y,c} \left(\int_{1}^{\tilde{y}} \frac{Q_{s=0,a}(X,c)}{\bar{G}^2_{s=0,a}(cjX)} dG_{s=0,a}(cjX) \right) P(C=cjY=y, X, S=0, A=a) P(Y(a)=yjX, S=0)
\end{aligned}$$

$$= \sum_{y,c} \left(\int_{1}^{\tilde{y}} \frac{Q_{s=1,a}(X,c)}{\bar{G}^2_{s=1,a}(cjX)} dG_{s=1,a}(cjX) \right) P(C = cjY = y, X, S = 1, A = a) P(Y(a) = yjX, S = 1)$$

$$= \mathsf{E}\left[\int_{1}^{\tilde{Y}} \frac{Q_{s=1,a}(X,c)}{\bar{G}^2_{s=1,a}(cjX)} dG_{s=1,a}(cjX) \Big| X, S = 1, A = a \right] \qquad \text{(by symmetry)}$$

where the second equality is by Lemma A.1 and the third by Lemma A.2 and Assumptions 2 and 5. □

A.8 Theorem 5

Theorem 5. *If Assumptions 1,2,5 hold, we have*

$$\mathsf{E}[\psi^{\mathsf{IPW},\tilde{Y}} \; j \; X] = \mathsf{E}[\psi^{\mathsf{IPW},\tilde{Y}}_{s=1} \quad \psi^{\mathsf{IPW},\tilde{Y}}_{s=0} \; j \; X] = 0. \tag{13}$$

Proof. We write the following $8a \; 2 \; f0, 1g$

$$\begin{aligned}
&\mathsf{E}\left[\frac{\mathbf{1}\, fS = 0, A = ag\, \tilde{Y}}{P(S = 0, A = ajX)} \Big| X \right] \\
&= \mathsf{E}\left[\tilde{Y} jX, S = 0, A = a \right] \\
&= \mathsf{E}\left[\min(Y, C) jX, S = 0, A = a \right] \\
&= \mathsf{E}\left[Y jX, S = 0, A = a, Y \quad C \right] P(Y \quad C jX, S = 0, A = a) \\
&\qquad + \mathsf{E}\left[C jX, S = 0, A = a, Y > C \right] P(Y > C jX, S = 0, A = a) \\
&= \mathsf{E}\left[Y jX, S = 1, A = a, Y \quad C \right] P(Y \quad C jX, S = 1, A = a) \qquad (24) \\
&\qquad + \mathsf{E}\left[C jX, S = 1, A = a, Y > C \right] P(Y > C jX, S = 1, A = a) \qquad \text{(Lemmas A.3 and A.4)} \\
&= \mathsf{E}\left[\frac{\mathbf{1}\, fS = 1, A = ag\, \tilde{Y}}{P(S = 1, A = ajX)} \Big| X \right] \qquad \text{(by symmetry)}
\end{aligned}$$

which immediately gives us

$$\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=0} jX \right] = \mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=1} jX \right]$$

by definition of $\psi^{\mathsf{IPW},\tilde{Y}}_{s}$ and we are done. □

A.9 Proposition 2

Proposition 2. *Consider the same setup in Theorems 4 and 5, with the only difference being we assume*

$$\mathsf{E}\left[Y(a) \; j \; X, S = 0 \right] = \mathsf{E}\left[Y(a) \; j \; X, S = 1 \right],$$

$8a \; 2 \; f0, 1g$*, instead of the stronger "ignorability of selection" in Assumption 2. Then, (11) and (13) are not true in general.*

Proof. In the first part of this proof, where we show that (13) is not true in general, we will consider the following distributions for the potential outcomes which satisfy the "mean exchangeability of the potential outcomes" (the

condition given in the statement of the proposition) but not the ignorability of selection condition in Assumption 2

$$
\begin{aligned}
P(Y(0) = n | S = 0) &= \begin{cases} 1/5 & n \in \{1,2,3,4,5\} \\ 0 & \text{otherwise} \end{cases} \\
P(Y(1) = n | S = 0) &= \begin{cases} 1/5 & n \in \{2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases} \\
P(Y(0) = n | S = 1) &= \begin{cases} 1/3 & n \in \{2,3,4\} \\ 0 & \text{otherwise} \end{cases} \\
P(Y(1) = n | S = 1) &= \begin{cases} 1/6 & n = 2 \\ 1/2 & n = 4 \\ 1/3 & n = 5 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{25}
$$

and the following *conditional* distributions for the censoring time

$$
\begin{aligned}
P(C = n | Y < 3.5) &= \begin{cases} 1 & n = 10 \\ 0 & \text{otherwise} \end{cases} \\
P(C = n | Y > 3.5) &= \begin{cases} 1 & n = 1/2 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{26}
$$

For simplicity, we further assume that

$$
\begin{aligned}
Y(0), Y(1) &\perp\!\!\!\perp X, A | S \\
C &\perp\!\!\!\perp X, S, A | Y
\end{aligned}
\tag{27}
$$

Note that (26) and (27) are stricter versions of no unobserved confounding (Assumption 1) and global censoring (Assumption 5). We then note, by (25), the following

$$
\begin{aligned}
\mathbb{E}[Y(0) | X, S = 0] &= 3 \\
\mathbb{E}[Y(1) | X, S = 0] &= 4 \\
\mathbb{E}[Y(0) | X, S = 1] &= 3 \\
\mathbb{E}[Y(1) | X, S = 1] &= 4
\end{aligned}
\tag{28}
$$

That is, mean exchangeability of the potential outcomes $\mathbb{E}[Y(a) | X, S = 0] = \mathbb{E}[Y(a) | X, S = 1]$, $\forall a \in \{0, 1\}$ holds. Note that, however, ignorability of selection (in Assumption 2) is violated (see (25)). Thanks to (27), we can simply calculate

$$
\begin{aligned}
P(Y \quad C | X, S = 0, A = 0) &= \frac{3}{5} \\
P(Y \quad C | X, S = 0, A = 1) &= \frac{2}{5} \\
P(Y \quad C | X, S = 1, A = 0) &= \frac{2}{3} \\
P(Y \quad C | X, S = 1, A = 1) &= \frac{1}{6}
\end{aligned}
\tag{29}
$$

since, *e.g.*, when $S = 0$ and $A = 0$, we have $Y \quad C$ if and only if $Y(0) \in \{1,2,3\}$, which happens with probability $3 \quad 1/5 = 3/5$. Next, we have

$$
\begin{aligned}
\mathbb{E}[Y | X, S = 0, A = 0, Y \quad C] &= \frac{1+2+3}{3} = 2 \\
\mathbb{E}[Y | X, S = 0, A = 1, Y \quad C] &= \frac{2+3}{2} = 5/2 \\
\mathbb{E}[Y | X, S = 1, A = 0, Y \quad C] &= \frac{2+3}{2} = 5/2 \\
\mathbb{E}[Y | X, S = 1, A = 1, Y \quad C] &= 2
\end{aligned}
\tag{30}
$$

and

$$\mathsf{E}\left[CjX, S, A, Y > C\right] = \frac{1}{2} \tag{31}$$

since $C = 1/2$ almost surely whenever $Y > C$. We are now ready to calculate

$$\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}jX\right] = \mathsf{E}\left[\left(\psi^{\mathsf{IPW},\tilde{Y}}_{s=1,a=1} \quad \psi^{\mathsf{IPW},\tilde{Y}}_{s=1,a=0}\right) \quad \left(\psi^{\mathsf{IPW},\tilde{Y}}_{s=0,a=1} \quad \psi^{\mathsf{IPW},\tilde{Y}}_{s=0,a=0}\right)\Big|X\right]$$

where

$$\psi^{\mathsf{IPW},\tilde{Y}}_{s,a} = \frac{\mathbf{1}\,fS = s, A = ag\tilde{Y}}{P(S = s, A = ajX)}$$

From (24), we have

$$\mathsf{E}\left[\psi^{\mathsf{IPW}}_{S,A}jX\right] = \mathsf{E}\left[YjX, S, A, Y \quad C\right]P(Y \quad CjX, S, A) + \mathsf{E}\left[CjX, S, A, Y > C\right]P(Y > CjX, S, A)$$

Plugging in the values calculated in (25), (26), (30), (31)

$$\begin{aligned}
\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=0,a=0}jX\right] &= \frac{3}{5} \quad 2 + \frac{2}{5} \quad \frac{1}{2} = \frac{7}{5} \\
\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=0,a=1}jX\right] &= \frac{2}{5} \quad \frac{5}{2} + \frac{3}{5} \quad \frac{1}{2} = \frac{13}{10} \\
\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=1,a=0}jX\right] &= \frac{2}{3} \quad \frac{5}{2} + \frac{1}{3} \quad \frac{1}{2} = \frac{11}{6} \\
\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}_{s=1,a=1}jX\right] &= \frac{1}{6} \quad 2 + \frac{5}{6} \quad \frac{1}{2} = \frac{3}{4}
\end{aligned} \tag{32}$$

We are done since

$$\mathsf{E}\left[\psi^{\mathsf{IPW},\tilde{Y}}jX\right] = \left(\frac{3}{4} \quad \frac{11}{6}\right) \quad \left(\frac{13}{10} \quad \frac{7}{5}\right) = \quad \frac{59}{60} \neq 0$$

Next, we show that (11) is not true in general. We keep the same setup above, only changing the marginal distributions of the potential outcomes to the following for simplicity

$$\begin{aligned}
P(Y(0) = njS = 0) &= \begin{cases} 1 & n = 0 \\ 0 & \text{otherwise} \end{cases} \\
P(Y(1) = njS = 0) &= \begin{cases} 1 & n = 2 \\ 0 & \text{otherwise} \end{cases} \\
P(Y(0) = njS = 1) &= \begin{cases} 1 & n = 0 \\ 0 & \text{otherwise} \end{cases} \\
P(Y(1) = njS = 1) &= \begin{cases} \frac{1}{2} & n \in f0, 4g \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{33}$$

Note that the mean exchangeability of potential outcomes again hold, but not the ignorability of selection in Assumption 2.

When $(s = 0, a = 0)$, $(s = 0, a = 1)$, or $(s = 1, a = 0)$, we have $C = 10$ almost surely, since $Y < 3.5$ almost surely and we never have censored outcomes (see (26)). We have $\tilde{Y} = Y$ and $\Delta = 1$ almost surely. Then for those values of $S = s$ and $A = a$, we can write, by (23)

$$\begin{aligned}
&\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s,a}jX\right] \\
=&\mathsf{E}\left[\frac{\mathbf{1}\,f\Delta = 1gY}{\bar{G}_{s,a}(YjX)} + \frac{\mathbf{1}\,f\Delta = 0gQ_{s,a}(X, C)}{\bar{G}_{s,a}(CjX)} \quad \int_{1}^{\tilde{Y}} \frac{Q_{s,a}(X, c)}{\bar{G}^2_{s,a}(cjX)} dG_{s,a}(cjX)\Big|X, S = s, A = a\right] \\
=&\mathsf{E}\left[\frac{Y}{\bar{G}_{s,a}(YjX)} \quad \int_{1}^{Y} \frac{Q_{s,a}(X, c)}{\bar{G}^2_{s,a}(cjX)} \delta(c \quad 10)\Big|X, S = s, A = a\right]
\end{aligned}$$

$$= \mathsf{E}\left[\frac{Y}{\bar{G}_{s,a}(Y|X)}\Big|X, S=s, A=a\right]$$

where δ is the Dirac delta function. The integral term disappears since $Y < 10$ almost surely. Since $\bar{G}_{s,a}(t|X) = 1$ for all $t < 10$, we can calculate the following

$$\begin{aligned}
\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0,a=0}|X\right] &= 0 \\
\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=0,a=1}|X\right] &= 2 \\
\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a=0}|X\right] &= 0
\end{aligned}$$

Next, we calculate $\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a=1}|X\right]$. Note that in this case, the observations are censored with probability $1/2$ (when $Y = 4$) and not censored with probability $1/2$ (when $Y = 0$). We then note the marginal (w.r.t. Y) survival function of the censoring time as

$$\bar{G}_{s=1,a=1}(n|X) = \begin{cases} 1 & n < \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \quad n < 10 \\ 0 & \text{otherwise} \end{cases}$$

And the density function as the sum of two Dirac delta functions

$$dG_{s=1,a=1}(n|X) = \frac{1}{2}\left(\delta(n \quad \frac{1}{2}) + \delta(n \quad 10)\right) \tag{34}$$

We can then write, also utilizing Lemma A.1 and (33)

$$\begin{aligned}
&\mathsf{E}\left[\psi^{\mathsf{CDR}}_{s=1,a=1}|X\right] \\
=&\mathsf{E}\left[\frac{\mathbf{1}\,\mathit{f}\Delta = 1\mathit{g}\,Y}{\bar{G}_{s=1,a=1}(Y|X)} + \frac{\mathbf{1}\,\mathit{f}\Delta = 0\mathit{g}\,Q_{s=1,a=1}(X,C)}{\bar{G}_{s=1,a=1}(C|X)} \quad \int_{1}^{\tilde{Y}} \frac{Q_{s=1,a=1}(X,c)}{\bar{G}^2_{s=1,a=1}(c|X)} dG_{s=1,a=1}(c|X)\Big|X, S=1, A=1\right] \\
=&\mathsf{E}\left[\frac{\mathbf{1}\,\mathit{f}\Delta = 1\mathit{g}\,Y}{\bar{G}_{s=1,a=1}(Y|X)} + \frac{\mathbf{1}\,\mathit{f}\Delta = 0\mathit{g}\,Q_{s=1,a=1}(X,C)}{\bar{G}_{s=1,a=1}(C|X)} \quad \int_{1}^{\tilde{Y}} \frac{Q_{s=1,a=1}(X,c)}{\bar{G}^2_{s=1,a=1}(c|X)} dG_{s=1,a=1}(c|X)\Big|X, Y=0, S=1, A=1\right] \\
&\qquad P(Y(1)=0|X, S=1) \\
+&\mathsf{E}\left[\frac{\mathbf{1}\,\mathit{f}\Delta = 1\mathit{g}\,Y}{\bar{G}_{s=1,a=1}(Y|X)} + \frac{\mathbf{1}\,\mathit{f}\Delta = 0\mathit{g}\,Q_{s=1,a=1}(X,C)}{\bar{G}_{s=1,a=1}(C|X)} \quad \int_{1}^{\tilde{Y}} \frac{Q_{s=1,a=1}(X,c)}{\bar{G}^2_{s=1,a=1}(c|X)} dG_{s=1,a=1}(c|X)\Big|X, Y=4, S=1, A=1\right] \\
&\qquad P(Y(1)=4|X, S=1) \\
=&\underbrace{\mathsf{E}\left[\frac{Y}{\bar{G}_{s=1,a=1}(Y|X)}\Big|X, Y=0, S=1, A=1\right]}_{=0} \quad \frac{1}{2} \qquad \text{(when } Y=0\text{, we have } C=10,\ \Delta=1, \tilde{Y}=0\text{)} \\
+&\mathsf{E}\left[\frac{Q_{s=1,a=1}(X,C)}{\bar{G}_{s=1,a=1}(C|X)} \quad \int_{1}^{\tilde{Y}} \frac{Q_{s=1,a=1}(X,c)}{\bar{G}^2_{s=1,a=1}(c|X)} dG_{s=1,a=1}(c|X)\Big|X, Y=4, S=1, A=1\right] \quad \frac{1}{2} \\
&\qquad \text{(when } Y=4\text{, we have } C=\tfrac{1}{2},\ \Delta=0, \tilde{Y}=\tfrac{1}{2}\text{)} \\
=&\left(\frac{Q_{s=1,a=1}(X,\frac{1}{2})}{\bar{G}_{s=1,a=1}(\frac{1}{2}|X)} \quad \int_{1}^{\frac{1}{2}} \frac{Q_{s=1,a=1}(X,c)}{\bar{G}^2_{s=1,a=1}(c|X)} dG_{s=1,a=1}(c|X)\right) \quad \frac{1}{2} \\
=&\left(\frac{Q_{s=1,a=1}(X,\frac{1}{2})}{\bar{G}_{s=1,a=1}(\frac{1}{2}|X)} \quad \frac{1}{2}\left(\frac{Q_{s=1,a=1}(X,\frac{1}{2})}{\bar{G}^2_{s=1,a=1}(\frac{1}{2}|X)}\right)\right) \quad \frac{1}{2} \qquad \text{(by (34))} \\
=&\left(\frac{4}{\frac{1}{2}} \quad \frac{1}{2}\frac{4}{\left(\frac{1}{2}\right)^2}\right) \quad \frac{1}{2}
\end{aligned}$$

=0

Note that $Q_{s=1,a=1}(X, \frac{1}{2}) = \mathsf{E}\left[YjX, Y > \frac{1}{2}, S = 1, A = 1\right] = 4$ since given $Y > \frac{1}{2}$, $Y = 4$ almost surely. We are done since

$$\mathsf{E}\left[\psi^{\mathsf{CDR}}jX\right] = \mathsf{E}\left[\left(\psi^{\mathsf{CDR}}_{s=1,a=1} \quad \psi^{\mathsf{CDR}}_{s=1,a=0}\right) \quad \left(\psi^{\mathsf{CDR}}_{s=0,a=1} \quad \psi^{\mathsf{CDR}}_{s=0,a=0}\right) \middle| X\right]$$
$$= (0 \quad 0) \quad (2 \quad 0) = \quad 2 \not\in 0$$

□

B DETAILS ON THE DATASETS USED IN THE EXPERIMENTS

B.1 IHDP Experiments

We use 10 covariates from the IHDP data for X, containing both continuous and categorical variables: `twin`, `b.head`, `preterm`, `momage`, `bw`, `b.marr`, `nnhealth`, `birth.o`, `momhisp`, `sex`.

B.1.1 Propensity Score of the Treatment

In the RCT cohort $S = 0$ we set $P(A = 1jX, S = 0) = 0.5$. That is, the treatment assignment is completely randomized. In the OS cohort, we have

$$P(A = 1jX, S = 1) = \text{sigmoid}(X^{\mid}\beta_{\text{prop}} + C)$$

where $C \in \{0.5, 0.75, 1, 1.25\}$ and $\beta_{\text{prop}}(\texttt{sex}) = 2 \quad C$ depending on the "strength" of confounding we want to induce. We set $\beta_{\text{prop}}(x) = 0$ for all $x \not\in \texttt{sex}$. That is, among the 10 covariates listed above, only the `sex` covariate has any influence on the treatment assignment. See the `prop_args` parameter in `exp_configs/ihdp/*.json` for the specific value of β_{prop} in each experimental setup.

The propensity score of the treatment is used to sample a binary treatment A_i for a patient with covariates x_i in the study s_i as

$$A_i \quad \texttt{Bernoulli}(P(A = 1jX = x_i, S = s_i))$$

B.1.2 Survival Functions of Time-to-Event and Censoring Time

Time-to-events Y and censoring times C are simulated using a CoxPH model (Cox, 1972) for their survival functions $\bar{F}_{s,a}(tjX)$ (4) and $\bar{G}_{s,a}(tjX)$ (3), which admits the general form

$$\overline{W}_0(t; \lambda, p)^{\exp(X^{\mid}\beta_{\text{Cox}})}$$

where

$$\overline{W}_0(t; \lambda, p) = \exp(\quad (\lambda t)^p) \qquad \lambda, p \in \mathsf{R}_+$$

Recall that the parameters λ, p, and $\beta_{\text{Cox}} \in \mathsf{R}^{10}$ are specified separately for each study $S \in \{0, 1\}$ and treatment group $A \in \{0, 1\}$ pair. The specific values for each parameter can be found in `exp_configs/ihdp/*.json` under the corresponding study key (RCT or OS).

To measure type-1 error (see setup #1 in Table 1), we use identical values for the parameters across the RCT and the OS. Identical parameters are also used in the experiments where we consider the violation of internal validity (see setups #4 and #5 in Table 1, and middle column in Figure 2). Only difference is that we conceal the `sex` covariate to induce unmeasured confounding. Finally, to simulate the violation of the external validity (see setups #2 and #3 in Table 1, and left column in Figure 2), we use different values for $\beta_{\text{Cox}}(\texttt{nnhealth})$ in the treatment $A = 1$ groups of the RCT $S = 0$ and the OS $S = 1$. We experiment with different magnitudes of difference for the parameter, denoted by $\Delta\beta_{\text{Cox}}$ in the main text (*e.g.*, see Table 1).

B.1.3 Estimation of the Nuisance Functions

Since we know the exact data-generating process, we are able to correctly specify the models for estimation.

For the selection $P(S = 1jX)$ and propensity $P(A = 1jX, S)$ scores, we fit logistic regression models. To limit the variance of the signals, we drop patients with extreme selection or propensity scores (*i.e.*, < 0.05 or > 0.95). For the survival functions $\bar{F}_{s,a}(tjX)$ (4) and $\bar{G}_{s,a}(tjX)$ (3), we fit a CoxPH model where the baseline survival function $\overline{W}_0(t; \lambda, p)$ is estimated via the Breslow estimator and β_{Cox} is estimated by fitting Cox's partial likelihood (Cox, 1972; Davidson-Pilon, 2019). Finally, we fit an XGboost model for the mean outcome regressors as part of the DR-$\tilde{Y}$ and DR-Y signals (see Appendix D).

B.2 WHI Experiments

WHI data is available to all researchers upon request in https://biolincc.nhlbi.nih.gov/studies/whi_ctos/. We start with 1121 features available at the baseline for both the RCT and the OS cohorts. After removing duplicate and highly correlated (Pearson coefficient > 0.95), we had 954 features. Finally, we transform to 350 principal components (PC), which capture 90% of the variance, and use those PCs as the set of X variables. THE PC transformation helps with the convergence and instability-related issues in the generalized linear models, specifically the Cox regression model. The estimation of the nuisance functions is done the same way as in the IHDP experiments (see Appendix B.1).

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 Testing the Doubly-Robustness of the CDR Signal

We test the CDR signal's doubly-robust property for setups #1,2,3 in Table 1 where internal validity holds.

We misspecify $\bar{F}_{S,A}(tjX)$ by setting the baseline survival function $\overline{W}_0(t; \lambda, p)$ to follow the law of a uniform random variable between the minimum and maximum time-to-event variables, and set $\beta_{\text{Cox}} = 0$. We use the correct (oracle) values for $\bar{G}_{S,A}(tjX)$ and $P(AjX, S)$, and name the resulting signal `CDR-MissF`.

We then misspecify $\bar{G}_{S,A}(tjX)$ and $P(AjX, S)$ and use the correct values for $\bar{F}_{S,A}(tjX)$, and name the resulting signal `CDR-MissGP`. The misspecification of $\bar{G}_{S,A}(tjX)$ is done similarly to that of $\bar{F}_{S,A}(tjX)$, by using the minimum and maximum censoring times. We also misspecify $P(AjX, S)$ by setting it to 0.5 in the OS $S = 1$.

We see in Table 3 that misspecified models maintain low type-1 error and higher power, corroborating the doubly-robustness of the CDR signal.

Table 3: Rejection rates over 40 runs for misspecified CDR signals. OS size is $n_1 = 2955$. Same setups in Table 1 are considered.

Setup #	**1**	**2**	**3**
Metric	**Type-1 error**	**Power**	**Power**
CDR-MissF	0	0.3	0.975
CDR-MissGP	0	0.35	0.975

C.2 Witness Function

As exposed in Hussain et al. (2023), an appealing feature of using MMR-based approach is that we can express the maximizer of M (Theorem 3) as

$$f \quad = \arg \sup_{f 2 F, jjfjj \quad 1} \left(\mathsf{E}\left[\psi f(X)\right]\right)^2$$

in closed form. This maximizer is referred to as the *witness function* and can be estimated as follows:

$$\hat{f} \quad = C \frac{1}{n} \sum_i \psi_i k(x_i, x)$$

where C is a constant such that $\int_{\mathcal{X}} f\ (x)dx = 1$.

The witness function reveals the regions of $\mathcal{X}$ where the "difference signal" ψ takes on larger values. As such, it can indicate the sub-populations where the CATE (see (1)) estimates from the RCT and OS are most discrepant.

We investigated the witness function using a fully synthetic dataset, where we control the "source covariates" of discrepancy and ensure limited correlation between them to facilitate the validation of the readings.

For both RCT $S = 0$ and OS $S = 1$, we generate covariates using one intercept and 10 independent normal variables with variances $\sigma^2 = 1$ and the following means

$$\mu_{s=0} = \big[0, 0, 0, 0, 0, 0, 0, 0, 1, 0\big]$$
$$\mu_{s=1} = \big[0, \quad 0.1, 0.4, 0, \quad 0.3, 0.15, 0, 0.4, 1, \quad 0.4\big]$$

For the OS, we generated the probability of treatment with a logistic regression with parameters β_{prop}:

$$P(A = 1 \mid X, S = 1) = \text{sigmoid}(X^{\intercal}\beta_{\text{prop}})$$
$$\text{where} \quad \beta_{\text{prop},s=1} = \big[\ 0.7, 0.4, \quad 0.2, 0.3, \quad 0.1, \quad 0.4, 0.2, 0.1, 0.4, \quad 0.8, \quad 0.75\big]$$

and for the RCT we have $P(A = 1 | X, S = 0) = 0.5$. We induce different CATE functions in the RCT and OS by using different parameter values for covariates X_8, X_9, and X_{10} in the CoxPH model for generating $Y(1)$.

$$\beta_{\text{Cox},S=0,Y(1)} = \big[0, 0.7, \quad 0.4, 0.5, 0.4, \quad 0.5, 0.6, \quad 0.4, 0.5, \quad 1.2, \quad 0.7\big]\,.$$
$$\beta_{\text{Cox},S=1,Y(1)} = \big[0, 0.7, \quad 0.4, 0.5, 0.4, \quad 0.5, 0.6, \quad 0.4, \quad 0.5, 1.2, 0.7\big]$$

The CoxPH model parameters for the potential outcome $Y(0)$ is set to be the same across studies. According to this data generating model, an effective witness function should enable the detection as X_8, X_9, and X_{10} as culprits for the discrepancy between RCT and OS.

In Figure 3, we scatter-plot (blue) the individual values of the difference signal ψ. The witness function at any point is then a weighted average of those values depending on the specific kernel function $k(\ ,\ )$. We visualize the linear fit (red) to the difference signal function ψ over X for some quick insight. We observe the strongest correlations for variables X_8, X_9, and X_{10}, as expected.

In Figure 4, we plot the witness function over each dimension individually (using the same values for the other dimensions, effectively excluding their effect on the result). We repeat the experiment with some additive noise on the corresponding covariate values to obtain uncertainty ranges. We observe that large X_8, X_9, and X_{10} result in high witness function values, pronouncing the difference in the corresponding β_{Cox} values across studies. The direction of growth indicates the sign of the discrepancy. These experiments confirm that the witness function can be used for identifying a sub-population that exhibits a stronger violation of the validity assumptions.

However, in practice, the covariates will be correlated, which can hamper the ability of the witness function to pinpoint the variables responsible for the discrepancy. Furthermore, if the discrepancy results from unmeasured confounders, our only hope would be to see some effect through features correlated with the unmeasured confounders. However, useful proxies for the unmeasured confounders should also alleviate the very issue of confounding; therefore, the witness functions' utility is fundamentally limited under unmeasured confounding.

C.3 Hypothesis Testing

After calculating the test statistic $\hat{\mathsf{M}}_n^2$, we follow Hussain et al. (2023) (see their Appendix E.2) to generate $B = 100$ samples from the null distribution H_0. Let $(w_{k1}, \ldots, w_{kn}) \quad \text{Multinom}(n, (\frac{1}{n}, \ldots, \frac{1}{n}))$. The k-th bootstrap sample of the null is given as

$$\hat{\mathsf{M}}_{n(k)}^2 = \frac{1}{n^2} \sum_{i,j \,\in\, 1,\ldots,n} (w_{ki} \quad 1)\hat{\psi}_i k(x_i, x_j)\hat{\psi}_j (w_{kj} \quad 1)$$

The p-value for the statistic is then calculated as

$$t_a = \frac{\Big[\sum_{k=1}^{B} \mathbf{1}(\hat{\mathsf{M}}_n^2 \quad \hat{\mathsf{M}}_{n(k)}^2)\Big] + 1}{B + 1}$$

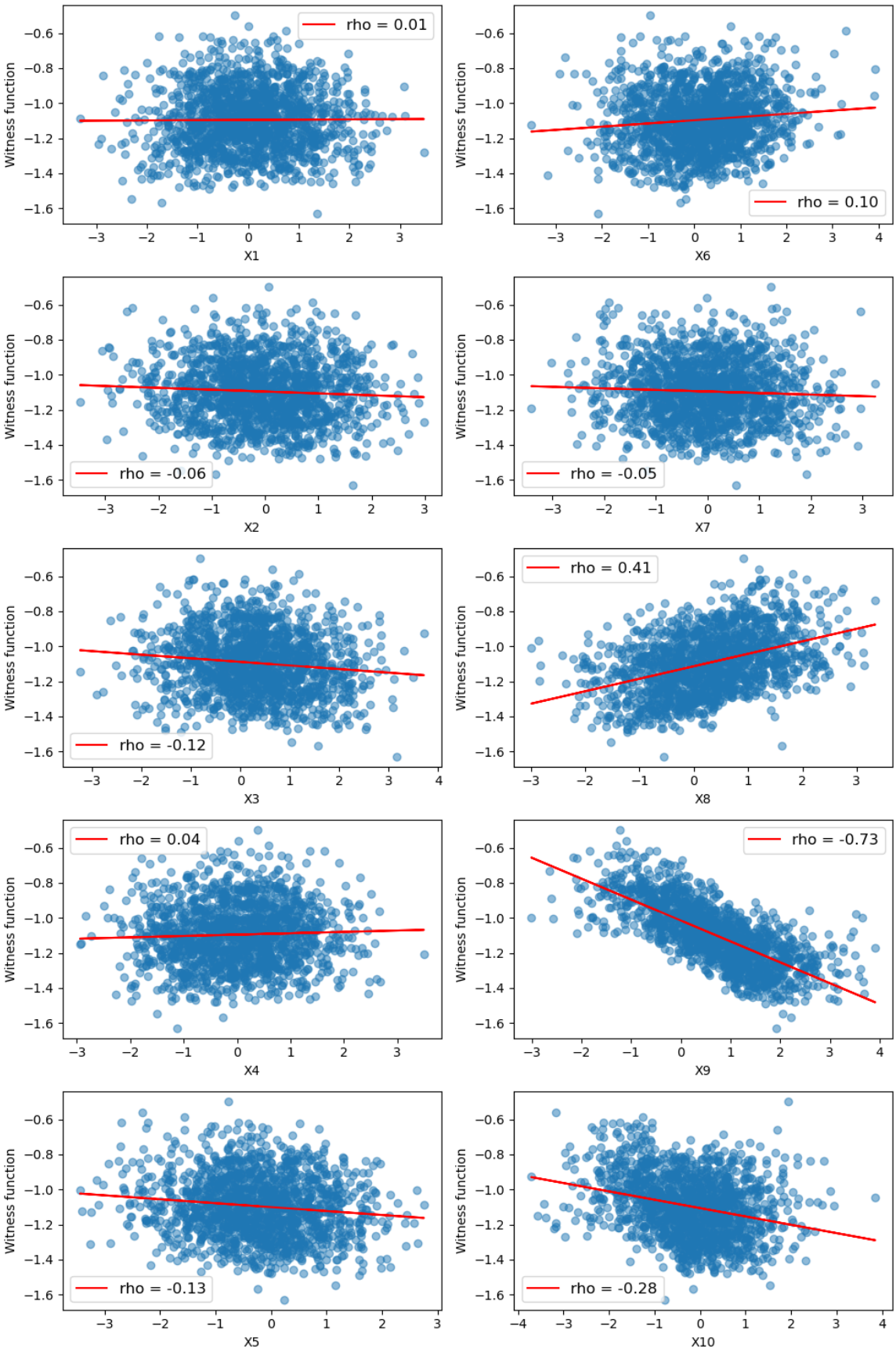

Figure 3: The difference signal ψ (blue) with a line fit to it (red) separately for each covariate.

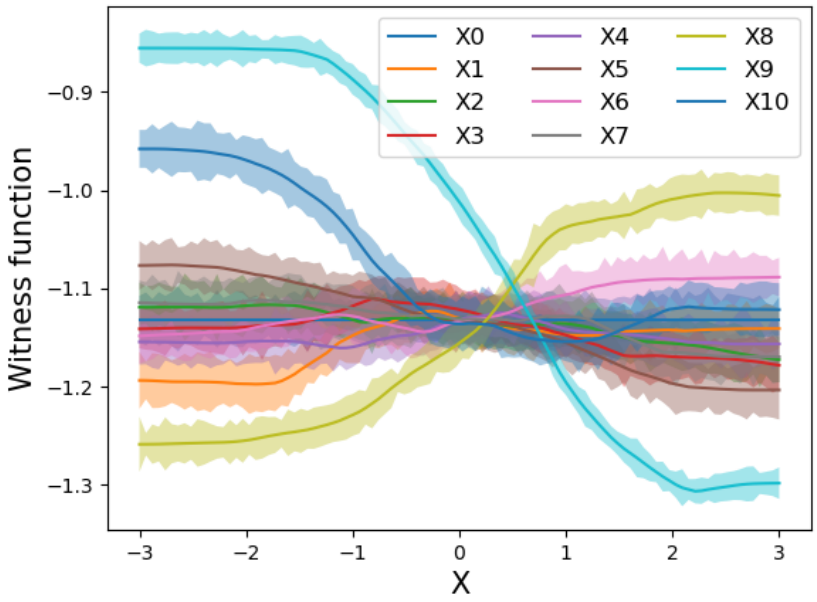

Figure 4: Witness function evaluated over each dimension individually.

If $t_a < 0.05$, we reject the null H_0 and accept it otherwise.

D BASELINE SIGNALS

We consider four signals without any component to model censoring that are used to construct the baseline tests in the ablation studies: IPW-Y, DR-Y, IPW-$\tilde{Y}$, and DR-$\tilde{Y}$. These signals are the standard inverse propensity weighting and regression-based signals in the literature. Let us start with the latter two. Note that $\psi^{\mathsf{IPW},\tilde{Y}}$ is already defined in (32). We define

$$\psi_{s,a}^{\mathsf{DR},\tilde{Y}} := \frac{1}{P(S=s|X)}\left(\frac{\mathbf{1}\{A=a\}\left(\tilde{Y} \quad \tilde{\mu}_{s,a}(X)\right)}{P(A=a|X,S=s)} + \tilde{\mu}_{s,a}(X)\right)$$
$$\psi_{s}^{\mathsf{DR},\tilde{Y}} := \psi_{s,a=1}^{\mathsf{DR},\tilde{Y}} \quad \psi_{s,a=0}^{\mathsf{DR},\tilde{Y}}$$
$$\psi^{\mathsf{DR},\tilde{Y}} := \psi_{s=1}^{\mathsf{DR},\tilde{Y}} \quad \psi_{s=0}^{\mathsf{DR},\tilde{Y}}$$

where

$$\tilde{\mu}_{s,a}(X) = \mathsf{E}\left[\tilde{Y}|X,S=s,A=a\right]$$

is the mean outcome function for the *imputed outcome* $\tilde{Y}$. $\psi^{\mathsf{IPW},Y}$ and $\psi^{\mathsf{DR},Y}$ are the conjugates of $\psi^{\mathsf{IPW},\tilde{Y}}$ and $\psi^{\mathsf{DR},\tilde{Y}}$, invoked only on the uncensored data ($\Delta = 1$). Precisely,

$$\psi_{s,a}^{\mathsf{DR},Y} := \frac{1}{P(S=s|X,\Delta=1)}\left(\frac{\mathbf{1}\{A=a\}\left(\tilde{Y} \quad \tilde{\mu}_{s,a}(X|\Delta=1)\right)}{P(A=a|X,S=s,\Delta=1)} + \tilde{\mu}_{s,a}(X|\Delta=1)\right)$$
$$\psi_{s}^{\mathsf{DR},Y} := \psi_{s,a=1}^{\mathsf{DR},Y} \quad \psi_{s,a=0}^{\mathsf{DR},Y}$$
$$\psi^{\mathsf{DR},Y} := \psi_{s=1}^{\mathsf{DR},Y} \quad \psi_{s=0}^{\mathsf{DR},Y}$$

where $\psi^{\mathsf{IPW},Y}$ is defined similarly and the signals are estimated using only the uncensored data.