Using a Surrogate with Heterogeneous Utility to Test for a Treatment Effect

Layla Parast[1], **Tianxi Cai**[2], **Lu Tian**[3]

[1]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX, USA

[2]Department of Biostatistics, Harvard University, Boston, MA, USA

[3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

Summary

The primary benefit of identifying a valid surrogate marker is the ability to use it in a future trial to test for a treatment effect with shorter follow-up time or less cost. However, previous work has demonstrated potential heterogeneity in the utility of a surrogate marker. When such heterogeneity exists, existing methods that use the surrogate to test for a treatment effect while ignoring this heterogeneity may lead to inaccurate conclusions about the treatment effect, particularly when the patient population in the new study has a different mix of characteristics than the study used to evaluate the utility of the surrogate marker. In this paper, we develop a novel test for a treatment effect using surrogate marker information that accounts for heterogeneity in the utility of the surrogate. We compare our testing procedure to a test that uses primary outcome information (gold standard) and a test that uses surrogate marker information, but ignores heterogeneity. We demonstrate the validity of our approach and derive the asymptotic properties of our estimator and variance estimates. Simulation studies examine the finite sample properties of our testing procedure and demonstrate when our proposed approach can outperform the testing approach that ignores heterogeneity. We illustrate our methods using data from an AIDS clinical trial to test for a treatment effect using CD4 count as a surrogate marker for RNA.

Keywords

1 | INTRODUCTION

There has been a substantial growth in clinical and methodological research on identifying and using valid surrogate markers in the past few decades. A valid surrogate marker is a biological measurement that can be used as a replacement for a primary outcome of interest in a clinical study. Many statistical methods have been proposed to evaluate and validate surrogate markers using a wide variety of innovative methodological approaches.[1,2,3,4,5] The primary benefit of identifying a valid surrogate marker is the ability to use it in a future trial to test for a treatment effect with less required follow-up time or less cost.

Correspondence: Layla Parast: parast@austin.utexas.edu.

For example, the U.S. Food and Drug Administration announced in 2020 that a surrogate marker that could be measured earlier than COVID-19 infection could be used to assess the vaccine efficacy in preventing infection,[6] thus potentially allowing for earlier identification of effective vaccines.

Several statistical methods have been proposed in recent years to assess the treatment effect on the primary outcome based on surrogate marker information. For example, Parast et al. (2019)[7] proposed a nonparametric approach to test for a treatment effect in a time-to-event outcome setting based on a surrogate marker measured at an earlier time point utilizing information about the relationship between the surrogate marker and primary outcome obtained from a prior study. Chen et al. (2020)[8] suggested a model-based approach that uses surrogate information to make interim decisions about whether to drop a treatment arm or stop a trial for futility. Price et al. (2018)[9] defined an optimal surrogate that optimally predicts a primary outcome and proposed super-learner and targeted super-learner based estimation procedures. Athey et al. (2019)[10] proposed to combine multiple surrogate markers to predict a long term outcome and estimate a treatment effect, and explicitly characterized the difference between the treatment effect estimated based on the primary outcome versus the surrogate combination.

Previous clinical and methodological work has demonstrated potential heterogeneity in the utility of a surrogate marker i.e. that a surrogate marker may be more useful (with respect to capturing the treatment effect on the primary outcome) for some subgroups than for others.[11] Parast et al. (2021)[12] offers a nonparametric estimation procedure and formal test for heterogeneity of surrogate utility with respect to a baseline covariate. When such heterogeneity exists, existing methods that use the surrogate to test for a treatment effect while ignoring this heterogeneity may lead to inaccurate conclusions about the treatment effect, particularly when the patient population in the current study has a different mix of characteristics than the prior study (used to evaluate the utility of the surrogate marker).

For example, in the simulation study in this paper, we examine a setting where the estimated treatment effect based on the primary outcome is 33.7 (standard error [SE] = 1.6); applying the testing approach of Parast et al. (2019)[7] which uses surrogate marker information but does not account for heterogeneity, the estimated treatment effect on the primary outcome is 39.2 (SE=3.5). The approach of Parast et al. (2019)[7] guarantees that the treatment effect based on the surrogate will be a lower bound for the true treatment effect on the primary outcome under certain conditions. However, these conditions may be violated when there is heterogeneity in the utility of the surrogate and thus leads to this type of situation where the estimated treatment effect using the surrogate is much higher than that using the primary outcome. Our approach that we propose in this paper which incorporates heterogeneity produces a treatment effect estimate that retains the lower bound property, with similar power to the treatment effect using the primary outcome. While we focus on heterogeneity with respect to a continuous baseline covariate, we provide a motivational example in Appendix A where there is heterogeneity with respect to a discrete covariate, gender. In this example, the surrogate marker is strong among males (explaining 99% of the treatment effect on the primary outcome) but weaker among females (explaining 67%). In a new study where the distribution of gender is 95% female and 5% male and the treatment effect on

the primary outcome is 38.95, using the surrogate marker and accounting for heterogeneity in surrogacy produces an estimated treatment effect on the primary outcome equal to 17.95 while ignoring heterogeneity produces an estimate of 44.5, again, failing to correctly provide a lower bound on the true treatment effect. In contrast, if we consider a future study where the distribution of gender is 5% female and 95% male, the treatment effect on the primary outcome is 74.05, while the treatment effect using the surrogate and accounting for heterogeneity is 71.05 versus not accounting for heterogeneity is 44.5, indicating a potential loss in power to detect a treatment effect when heterogeneity is ignored.

In this paper, we develop a novel test for a treatment effect using surrogate marker information that accounts for heterogeneity in the utility of the surrogate. We compare our testing procedure to a test that uses primary outcome information only (gold standard) and a test that uses surrogate marker information, but ignores heterogeneity. We demonstrate the validity of our testing procedure and derive the asymptotic properties of our estimator and variance estimates. A simulation study is used to examine the finite sample properties of our testing procedure and demonstrate when our proposed approach can outperform the testing approach that ignores heterogeneity. In particular, we demonstrate examples where the test of Parast et al. (2019)[7] provides an incorrect estimate with respect to the treatment effect. We illustrate our approach using data from an AIDS clinical trial to test for a treatment effect using CD4 count as a surrogate marker for plasma HIV-1 RNA.

2 | TESTING PROCEDURE

2.1 | Notation and Setting

We focus on a setting where we are currently conducting a study to examine the effect of a treatment on a primary outcome of interest, denoted by Y, and we additionally have data available from a prior study. We assume that this prior study was used to examine the strength of the surrogate, denoted by S, and heterogeneity in the utility of the surrogate, and has measurements of both Y and S of the current study. Let Z denote the treatment indicators where treatment is randomized and $Z \in \{0, 1\}$ (i.e., treatment vs. control), and W denote a baseline covariate such that S has been shown to have heterogeneous utility with respect to this covariate. Without loss of generality, we take W to be continuous; all proposed procedures can easily accommodate a discrete W as well. We focus on a setting with heterogeneity with respect to a single baseline covariate W; in Section 3.3, we discuss an extension to multiple W. In addition, we assume we are in a setting where either S is measured earlier than Y or S is measured at the same time as Y but is less expensive, invasive or burdensome, and there is no censoring or missing data. Throughout this paper, we quantify surrogate strength/utility using the quantity: the proportion of treatment effect on the primary outcome explained by the treatment effect on the surrogate marker.[13,3,5]

We use potential outcomes notation where each person has a potential $\left\{Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\right\}$ where $Y^{(g)}$ is the outcome when $Z = g$ and $S^{(g)}$ is the surrogate when $Z = g$. Observed data from the current study is denoted as and consists of $\mathscr{D} = \{(Y_{gi}, S_{gi}, W_{gi}), i = 1, \ldots, n_g; g = 0, 1\}$, where n_g denotes the number of individuals in treatment group g.

The goal in the current study is to test for a treatment effect on the primary outcome quantified as

$$H_0: \Delta \equiv E\left(Y^{(1)} - Y^{(0)}\right) = E\left(Y^{(1)}\right) - E\left(Y^{(0)}\right) = 0.$$

Our aim is to leverage information from the *prior* study to test H_0 using surrogate marker information in order to reduce study follow-up time, costs, and/or participant burden, i.e., making inference on Δ without using $\{Y_{gi}, i = 1, \cdots, 1, n_g; g = 0, 1\}$. We use a superscript p to denote "prior" when referring to data or quantities from the prior study. For example, we denote observed data from the prior study by $\mathscr{D}^p = \{(Y_{gi}^p, S_{gi}^p, W_{gi}^p, i = 1, \ldots, n_g^p, g = 0, 1\}$, where n_g^p is the sample size of treatment group g.

2.2 | Assumptions

Given that our setting rests on the existence of a valid surrogate marker, we first define S to be a valid surrogate marker for Y if the following conditions hold:

(C1) $E\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)$ is a monotone function of s;

(C2) $P\left(S^{(1)} > s \mid W = w\right) \geq P\left(S^{(0)} > s \mid W = w\right)$ for all s and w;

(C3) $E\left(Y^{(1)} \mid S^{(1)} = s, W = w\right) \geq E\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)$ for all s and w.

(C4) A large proportion of the treatment effect on the primary outcome can be explained by the treatment effect on the surrogate marker for all w.

Assumptions (C1)-(C3) are parallel to those required in Wang and Taylor (2002)[3] and Parast et al. (2017)[14] and protect against the surrogate paradox situation.[15] Assumption (C1) implies that the surrogate marker is either "positively" or "negatively" related to the time of the primary outcome, (C2) implies that there is a positive treatment effect on the surrogate marker, and (C3) implies that there is a non-negative residual treatment effect beyond that on the surrogate marker. Assumptions (C1-C3) together guarantee that $E\left(Y^{(1)} \mid W = w\right) \geq E\left(Y^{(0)} \mid W = w\right)$, for all w in the support of W (see Appendix B). Lastly, (C4) states that the proportion of the treatment effect explained by the surrogate marker must be large and guarantees the strength of the surrogate marker of interest for all individuals in the study. While this is somewhat vague, there is no agreed upon value that signifies a "large" proportion, though previous work has tended to view values of 0.6-0.75 or higher as large.[16,13,17] If the existing heterogeneity is such that the surrogate is strong for some w and weak for other w, it should *not* be used as a replacement of the primary outcome for all individuals in a future study. Instead, one may consider using the surrogate as a replacement only among those with a W where the surrogate is strong; we discuss this further in the Discussion.

In order to ensure that the proposed test statistic to be described in Section 2.3, has a reasonable interpretation with respect to Δ, we also require:

(C5) $E\left(Y^{(0)} \mid S^{(0)} = s, W = w\right) = E\left(Y^{(0p)} \mid S^{(0p)} = s, W^p = w\right)$ for all s and w;

(C6) $E\left(Y^{(0p)} \mid S^{(0p)} = s, W^p = w\right)$ is estimable for any $(s, w) \in \Omega_J$, where Ω_J is the common compact support for both $\left(S^{(g)}, W^{(g)}\right)$ in $g = 0, 1$.

Assumption (C5) implies that in the control groups, the current study and the prior study share the same conditional expectation for Y given S and W. This assumption is reasonable when, for example, the control condition in both studies are the same, such as "usual care." Importantly, such an assumption is not required to hold for the treatment groups and it relaxes the requirement that the distribution of Y conditional on S be transportable from the prior to current study. Even so, this assumption is admittedly very strong and needs to be carefully considered before using this approach; however, *any* testing procedure that attempts to borrow information from a prior study to test a hypothesis in a future study is going to require some type of strong transportability assumption. If there is reason to believe that such transportability between studies is not appropriate, then the prior study should not be considered for informing the future study. Assumption (C6) ensures that we can approximate $E\left(Y^0 \mid S^0 = s, W^0 = w\right)$ for all observed pairs of $S^{(g)}$ and $W^{(g)}, g = 0, 1$ in the current study. We discuss robustness to these assumptions as well as additional assumptions needed for a causal interpretation in Appendix B.

2.3 | Proposed Testing Procedure

Recall that our aim is to take advantage of information from the prior study to test H_0 using surrogate marker information such that this test accounts for known heterogeneity in the utility of the surrogate marker. To achieve this goal we note that Δ can be expressed as:

$$
\begin{aligned}
\Delta &= E\left(Y^{(1)}\right) - E\left(Y^{(0)}\right) = \int \Delta(w) dF_W(w) \\
&= \int \left[\int \mu_1(s, w) dF^{(1)}(s \mid w)\right] dF_W(w) - \int \left[\int \mu_0(s, w) dF^{(0)}(s \mid w)\right] dF_W(w)
\end{aligned}
\tag{1}
$$

where $\mu_g\left(s, w\right) \equiv E\left(Y^{(g)} \mid S^{(g)} = s, W = w\right), F^{(g)}\left(s \mid w\right) \equiv F_{S^{(g)} \mid W}\left(s \mid w\right)$ is the conditional cumulative distribution function of $S^{(g)}$ given $W = w$, and $F_W(w)$ is the cumulative distribution of W. In expressing Δ as (1), we have simply used a conditional expectation to incorporate S and W into our expression. By expressing Δ in this way, this motivates the following *earlier* treatment effect definition:

$$
\Delta_H = \int \left[\int \mu_0(s, w) dF^{(1)}(s \mid w)\right] dF_W(w) - \int \left[\int \mu_0(s, w) dF^{(0)}(s \mid w)\right] dF_W(w)
\tag{2}
$$

$$
= \int \mu_0^p(s, w) dF^{(1)}(s, w) - \int \mu_0^p(s, w) dF^{(0)}(s, w)
\tag{3}
$$

where $F^{(g)}(s, w)$ is the cumulative distribution function of $\left(S^{(g)}, W\right)$ in the current study. The only change in going from (1) to (2), is that we have replaced $\mu_1(s, w)$ with $\mu_0(s, w)$ in the first term which will ensure that this quantity provides a lower bound on the treatment effect. In the second equality, (3), we replace $\mu_0(s, w)$ with $\mu_0^p(s, w)$ which follows from Assumption (C5). The expression (3) is now a quantity that only involves $\mu_0^p(s, w)$ which is the conditional risk in the prior study, and the distribution of S and W in the current study. Importantly, the expression does not involve Y from the current study at all. In practice, $\mu_0^p(s, w)$ is unknown and must be replaced with an estimate, $\hat{\mu}_0^p(s, w)$, which we describe in Section 3.1. Because of this, we define the following *earlier* average treatment effect quantity, where the $\tilde{\ }$ notation makes the dependence on information from the prior study explicit:

$$\widetilde{\Delta}_H = \int \hat{\mu}_0^p(s, w) dF^{(1)}(s, w) - \int \hat{\mu}_0^p(s, w) dF^{(0)}(s, w) = E\left\{\hat{\mu}_0^p\left(S^{(1)}, W\right) - \hat{\mu}_0^p\left(S^{(0)}, W\right) \mid \mathscr{D}^p\right\}.$$

This quantity, $\widetilde{\Delta}_H$, measures the treatment effect on a transformation of the surrogate marker and baseline covariate, i.e., the difference between $\hat{\mu}_0^p\left(S^{(1)}, W\right)$ and $\hat{\mu}_0^p\left(S^{(0)}, W\right)$. First, due to randomization, W has the same distribution between two treatment groups and $\widetilde{\Delta}_H$ has an appealing causal interpretation reflecting the treatment effect on the surrogate marker. Second, $\widetilde{\Delta}_H$ represents the part of the treatment effect on the primary outcome explained by the surrogate marker and an approximation to Δ_H, which is the quantity of our primary interest. Under the null hypothesis of no average treatment effect on the primary outcome, there will also be no average treatment effect in any subgroup of patients with $W = w$ (see Appendix B). Under the null, Assumptions (C1)-(C3) imply that $S^{(1)} \mid W = w$ has the same distribution as $S^{(0)} \mid W = w$ for all w in the support of W, and thus, $\widetilde{\Delta}_H = 0$. Therefore, we may formally define our test statistic for H_0 based on the early average treatment effect as $Z_H = \sqrt{n}\, \widehat{\Delta}_H / \hat{\sigma}_H$, where $\widehat{\Delta}_H$ is a root-n consistent estimate of $\widetilde{\Delta}_H$ and $\hat{\sigma}_H^2$ is the estimated variance of $\sqrt{n}\left(\widehat{\Delta}_H - \widetilde{\Delta}_H\right)$. We reject H_0 when $|Z_H|$ is large. In Section 3, we propose robust procedures to construct $\hat{\Delta}_H$ and $\hat{\sigma}_H$. Obviously, this is a valid test for both the null $H_{0H}\colon \widetilde{\Delta}_H = 0$ and the null $H_0\colon \Delta = 0$.

One important merit of constructing the test statistic based on an estimator of $\widetilde{\Delta}_H$ is that this earlier average treatment effect is smaller than if we used the true conditional expectations within each treatment group in probability. That is, $P\left(\widetilde{\Delta}_H \leq \Delta\right) \approx 1$ and thus, $\widetilde{\Delta}_H$ is a conservative measure of the average treatment effect, Δ. Importantly, this early treatment effect and associated test account for heterogeneity in the utility of the surrogate by explicitly utilizing a condition mean function that depends on W. In the following section we describe other tests that may be considered; in our numerical studies, we compare our approach with these alternatives.

2.4 | Alternative Testing Approaches

We consider two alternative tests that would be reasonable options for testing H_0 in this setting. The first quite obvious approach is simply to assume the primary outcome is

measured in the current study and use primary outcome information to estimate Δ and conduct a t-test of $H_0: \Delta = 0$. This reflects the gold standard as it directly tests the hypothesis we are interested in. Importantly though, the whole point of this setting is to provide a way to *not* have to measure the primary outcome. We include this option so that we can compare to this gold standard.

The second alternative test we examine is one which uses information from the prior study about the relationship between the surrogate and the primary outcome, but does not account for heterogeneity. This test is an extension of a test proposed in Parast et al. (2019)[7] which was developed for the time-to-event outcome setting. Our description of it here, for a non-survival setting, is new and will be useful in practice for those analyzing a non-survival study in a setting with no heterogeneity in the utility of the surrogate. Similar to our proposed test, but without regard for W, we note that $\Delta = \int \mu_1(s) dF^{(1)}(s) - \int \mu_0(s) dF^{(0)}(s)$ where $\mu_g(s) = E(Y^{(g)} \mid S^{(g)})$ which motivates the following *earlier* treatment effect definition:

$$\Delta_P = \int \mu_0(s) dF^{(1)}(s) - \int \mu_0(s) dF^{(0)}(s) = \int \mu_0^p(s) dF^{(1)}(s) - \int \mu_0^p(s) dF^{(0)}(s)$$

where $\mu_0^p(s) \equiv E(Y^{(0p)} = y \mid S^{(0p)} = s)$. Since $\mu_0^p(s)$ is unknown, we approximate Δ_P with

$$\widetilde{\Delta}_P = \int \hat{\mu}_0(s) dF^{(1)}(s) - \int \hat{\mu}_0(s) dF^{(0)}(s) = \int \hat{\mu}_0^p(s) dF^{(1)}(s) - \int \hat{\mu}_0^p(s) dF^{(0)}(s).$$

where $\hat{\mu}_P^0(s)$ is a consistent estimator of $\mu_0^p(s)$. As with the proposed test, this early treatment effect quantity replaces $\mu_g(s)$ with $\hat{\mu}_0(s)$ for both treatment groups and will ensure it is a lower bound on the Δ under certain conditions. This test, however, requires the assumption that $\hat{\mu}_0^p(s) \approx \mu_0^p(s) = \mu_0(s)$ i.e., that this conditional expectation in the control group is the same in the current study as the prior study. It is important to note that this assumption may not hold when there is heterogeneity in the utility of the surrogate marker. To test $H_0: \Delta = 0$, we instead test $H_{0P}: \widetilde{\Delta}_P = 0$ and define the test statistic for H_{0P} based on the early treatment effect as $Z_P = \sqrt{n} \, \widehat{\Delta}_P / \hat{\sigma}_P$, where $\widehat{\Delta}_P$ is a root-n consistent estimate of $\widetilde{\Delta}_P$ and $\hat{\sigma}_P^2$ is the estimated variance of $\sqrt{n}(\widehat{\Delta}_P - \widetilde{\Delta}_P)$. We reject $H_{0P}($ and $H_0)$ when $|Z_P|$ is large.

In Appendix C, we discuss estimation and testing for Δ using the primary outcome, propose estimation procedures to obtain $\widehat{\Delta}_P$ and $\hat{\sigma}_P$, and discuss why we do not consider directly testing the surrogate. Intuitively, we would expect that both our proposed test and this test based on $\widetilde{\Delta}_P$ should work well when there is no heterogeneity. When there is heterogeneity, we expect that the test based on $\widetilde{\Delta}_P$ (or even Δ_P) could lead to erroneous conclusions about the treatment effect and/or have less power than the proposed test.

3 | ESTIMATION AND INFERENCE

3.1 | Estimation of Proposed $\widetilde{\Delta}_H$

For our proposed testing procedure, we first define

$$\hat{\mu}_0^p(s, w) = \frac{\sum_{i=1}^{n_0^p} K_{h_2}(S_{0i}^p - s) K_{h_3}(W_{0i}^p - w) Y_{0i}^p}{\sum_{i=1}^{n_0^p} K_{h_2}(S_{0i}^p - s) K_{h_3}(W_{0i}^p - w)}, \text{ and}$$

$$\hat{m}_g(w; \mu(\,\cdot\,,\,\cdot\,)) = \frac{\sum_{i=1}^{n_g} K_{h_g}(W_{gi} - w) \mu(S_{gi}, W_{gi})}{\sum_{i=1}^{n_g} K_{h_g}(W_{gi} - w)},$$

as nonparametric smoothed estimators of the conditional expectation of $Y^{(0)}$ given $\left(S^{(0)}, W\right) = \left(s, w\right)$ in the prior study, and the conditional expectation of $\mu\left(S^{(g)}, W\right)$ given $W = w$ and a bivariate function $\mu(\,\cdot\,,\,\cdot\,)$ in the current study, respectively. Here, $K_h(\,\cdot\,) = K(\,\cdot\,/h)/h$, $K(\,\cdot\,)$ is a smooth symmetric density function with finite support, h_0, h_1, h_2, h_3 are specified bandwidths which may be data dependent, and n_0^p denotes the sample size of group $Z = 0$ in the prior study. We utilize undersmoothing and select all bandwidths throughout to be of order $O(n^{-\epsilon}), \epsilon \in (1/4, 1/2)$, to eliminate the asymptotic bias, where $n = n_1 + n_0$ in an effort to avoid a need for bias correction in subsequent statistical inference.

A very straightforward estimate of $\widetilde{\Delta}_H$ would be

$$n_1^{-1} \sum_{i=1}^{n_1} \hat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - n_0^{-1} \sum_{i=1}^{n_0} \hat{\mu}_0^{(p)}(S_{0i}, W_{0i}) \tag{4}$$

which simply takes our estimated conditional mean function from the prior study and applies it to data in the current study. However, it is possible for us to improve upon this estimator in terms of efficiency. To do this, we note that

$$\begin{aligned} \widetilde{\Delta}_H &= E\left[E\left(\hat{\mu}_0^p\left(S^{(1)}, W\right) \mid W \right) \right] - E\left[E\left(\hat{\mu}_0^p\left(S^{(0)}, W\right) \mid W \right) \right] \\ &\approx E[\hat{m}_1(W; \hat{\mu}_0^p)] - E[\hat{m}_0(W; \hat{\mu}_0^p)], \end{aligned}$$

and thus we now consider an estimate of $\widetilde{\Delta}_H$ as

$$n_1^{-1} \sum_{i=1}^{n_1} \hat{m}_1(W_{1i}; \hat{\mu}_0^p) - n_0^{-1} \sum_{i=1}^{n_0} \hat{m}_0(W_{0i}; \hat{\mu}_0^p), \tag{5}$$

which is asymptotically equivalent to (4). Note that this estimate only uses $S^{(g)}$ and W data from the current study (no Y data from the current study) and $\hat{\mu}_0^p(s, w)$, which in turns depends on $S^{(0p)}, W^p, Y^{(0p)}$ data in group $Z = 0$ from the previous study.

While either (4) or (5) would be consistent estimates of $\widetilde{\Delta}_H$, we utilize the fact that the distributions of W from the two treatment arms are identical due to randomization and construct the estimator:

$$\widehat{\Delta}_H = \frac{1}{n_1 + n_0} \left\{ \left[\sum_{i=1}^{n_0} \widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) + \sum_{i=1}^{n_1} \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) \right] - \left[\sum_{i=1}^{n_0} \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) + \sum_{i=1}^{n_1} \widehat{m}_0(W_{1i}; \widehat{\mu}_0^p) \right] \right\}.$$

(6)

We show in Appendix D that (6) improves upon the efficiency of (5). Essentially, $\widehat{\Delta}_H$ is equivalent to an augmented version of the simple estimator (described below), taking advantage of the independence of W and treatment, since treatment was randomized.

In Appendix D we show that conditional on $\widehat{\mu}_0^p(\cdot, \cdot)$, $\widehat{\Delta}_H$ is a consistent estimate of $\widetilde{\Delta}_H$, and that $\sqrt{n}\left\{ \widehat{\Delta}_H - \widetilde{\Delta}_H \right\}$ weakly converges to a mean zero normal distribution as $n \to \infty$. A consistent estimate of the conditional variance of $\hat{\Delta}_H$ given the prior study, σ_H^2, can be obtained as

$$\hat{\sigma}_H^2 = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \left(\widetilde{S}_{1i} - \pi_0 \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widehat{\Delta}_H \right)^2$$
$$+ \frac{1}{n_0^2} \sum_{i=1}^{n_0} \left(\widetilde{S}_{0i} - \pi_0 \widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widehat{\Delta}_H \right)^2$$

where $\pi_g = n_g/n$ and $\widetilde{S}_{gi} = \mu_0^{(p)}(S_{gi}, W_{gi})$. Our testing procedure uses the test statistic $Z_H = \widehat{\Delta}_H / \hat{\sigma}_H$ and rejects the null hypothesis when $|Z_H| > \Phi^{-1}(1 - \alpha/2)$. As $n_{0p} \to \infty$, $\widetilde{\Delta}_H - \Delta_H = o_p(1)$ and $\widetilde{\Delta}_H$ can be viewed as a consistent estimator of Δ_H. More importantly, under Assumptions (C1), (C2), (C3) and (C5), $P\left(\widetilde{\Delta}_H \leq \Delta \right) \to 1$ as $n \to \infty$, indicating that the test for $\widetilde{\Delta}_H = 0$ is a valid test for $\Delta = 0$ with probability approaching 1 as the sample size of the prior study increases to infinity.

Remark. *The efficiency of the simple estimator*

$$n_1^{-1} \sum_{i=1}^{n_1} \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - n_0^{-1} \sum_{i=1}^{n_0} \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) \approx n_1^{-1} \sum_{i=1}^{n_1} \widehat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - n_0^{-1} \sum_{i=1}^{n_0} \widehat{\mu}_0^{(p)}(S_{0i}, W_{0i}),$$

can be improved by considering the fact that $E[m(W_{1i}; \widehat{\mu}_0^p)] = E[m(W_{0i}; \widehat{\mu}_0^p)]$ for any transformation $m(\cdot)$ due to randomization. Specifically, one may consider a new class of consistent estimators indexed by $m(\cdot): R \to R$,

$$\left\{ n_1^{-1} \sum_{i=1}^{n_1} \left[\widehat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - m(W_{1i}; \widehat{\mu}_0^p) \right] - n_0^{-1} \sum_{i=1}^{n_0} \left[\widehat{\mu}_0^{(p)}(S_{0i}, W_{0i}) - m(W_{0i}; \widehat{\mu}_0^p) \right] \right\}.$$

The optimal choice of $m(\,\cdot\,)$ minimizing the asymptotic variance is

$$m_{opt}(w) = \pi_0 E\big(\widehat{\mu}_0^{(p)}(S_1, w) \mid W_1 = w\big) + \pi_1 E\big(\widehat{\mu}_0^{(p)}(S_0, w) \mid W_0 = w\big).$$

In practice, $m_0(w)$ can be consistently estimated by $\widehat{m}_{opt}(w) = \pi_0 \widehat{m}_1\big(w; \widehat{\mu}_0^{(p)}\big) + \pi_1 \widehat{m}_0\big(w; \widehat{\mu}_0^{(p)}\big)$. Denote the resulting estimator of $\widetilde{\Delta}_H$ by

$$\widehat{\Delta}_H^{AUG} = n_1^{-1} \sum_{i=1}^{n_1} \left[\widehat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - \widehat{m}_{opt}(W_{1i}; \widehat{\mu}_0^p) \right] - n_0^{-1} \sum_{i=1}^{n_0} \left[\widehat{\mu}_0^{(p)}(S_{0i}, W_{0i}) - \widehat{m}_{opt}(W_{0i}; \widehat{\mu}_0^p) \right].$$

In Appendix D *we show that conditional on $\widehat{\mu}_0^{(p)}\big(\,\cdot\,,\,\cdot\,\big)$, $\widehat{\Delta}_H^{AUG}$ is a consistent estimate of $\widetilde{\Delta}_H$ and that $\sqrt{n}\Big(\widehat{\Delta}_H^{AUG} - \widetilde{\Delta}_H \Big)$ weakly converges to a mean zero normal distribution as $n \to \infty$. The conditional variance of $\widehat{\Delta}_H^{AUG} \mid \widehat{\mu}_0^{(p)}\big(\,\cdot\,,\,\cdot\,\big), \sigma_{AUG}^2$, can be consistently estimated by*

$$\widehat{\sigma}_{AUG}^2 = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \left[\widehat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) \right]^2 + \frac{1}{n_0^2} \sum_{i=1}^{n_0} \left[\widehat{\mu}_0^{(p)}(S_{0i}, W_{0i}) - \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) \right]^2$$
$$+ \frac{\pi_1^2}{n_1^2} \sum_{i=1}^{n_1} \left[\widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - \widehat{m}_0(W_{1i}; \widehat{\mu}_0^p) - \widehat{\Delta}_H \right]^2 + \frac{\pi_0^2}{n_0^2} \sum_{i=1}^{n_0} \left[\widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) - \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) - \widehat{\Delta}_H \right]^2.$$

In Appendix D, *we show that $\widehat{\Delta}_H^{AUG}$ is asymptotically equivalent to our proposed $\widehat{\Delta}_H$ and $\widehat{\sigma}_H / \widehat{\sigma}_{AUG} = 1 + o_p(1)$.*

3.2 | Inference

To construct a confidence interval for $\widetilde{\Delta}_H$ we use our estimated variance $\widehat{\sigma}_H^2$ and define a $100(1-\alpha)\%$ confidence interval as $\widehat{\Delta}_H \pm Z_{1-\alpha/2}\widehat{\sigma}_H$. We examine the empirical performance of our proposed estimation procedure, variance estimation, confidence interval construction, and testing procedure in Section 4.

It is important to note that we consider the prior study, the study from which we estimate the conditional mean function, $\widehat{\mu}_0^p(s, w)$, as fixed. This is a reasonable assumption given that in practice, there is truly some previously conducted prior study which one is using to inform testing in the current study. However, one could argue that this prior study should be considered random and that all inference should be derived as such. In such a case, the estimation of our point estimate $\widehat{\Delta}_H$ would remain the same but the standard estimation and confidence interval construction would be more complex.

3.3 | Multiple Baseline Covariates

While in this paper we focus only on heterogeneity with respect to a single baseline covariate, it may be the case that there is heterogeneity with respect to multiple baseline covariates. In such a case, one still can consider a straightforward estimator for the treatment effect using surrogate marker and baseline covariates:

$$n_1^{-1} \sum_{i=1}^{n_1} \widehat{\mu}_{0m}^{(p)}(S_{1i}, \mathbf{W}_{1i}) - n_0^{-1} \sum_{i=1}^{n_0} \widehat{\mu}_{0m}^{(p)}(S_{0i}, \mathbf{W}_{0i})$$

where $\widehat{\mu}_{0m}^{(p)}(s, \mathbf{w})$ is an estimator of $\mu_0(s, \mathbf{w}) \equiv E(Y^{(0)} \mid S^{(0)} = s, \mathbf{W} = \mathbf{w})$ and $\mathbf{W}$ is a baseline covariate vector of interest (including an intercept term, with a slight abuse of notation). The difficulty is that fully nonparametric estimation of $\mu_0(s, \mathbf{w})$ will likely be infeasible for practical sample sizes with a vector $\mathbf{W}$ of moderate dimension, e.g., 3. In such a case, one may be willing to consider a parametric or semi-parametric model. For example, an estimator can be obtained based on a simple regression model $\mu_0(s, \mathbf{w}) = g_Y(\beta_0 s + \boldsymbol{\beta}_1' \mathbf{w})$, where $g_Y(\cdot)$ is a known, strictly increasing link function and β_0 and $\boldsymbol{\beta}_1$ are unknown regression coefficients to be estimated based on the prior study. Alternatively, one could consider a more flexible varying coefficient model for $\mu_0^p(s, \mathbf{w})$ such as $\mu_0(s, \mathbf{w}) = g_Y\{\mathbf{B}(s)' \mathbf{w}\}$, where $\mathbf{B}(s) = \{\beta_1(s), \beta_2(s), \ldots, \beta_L(s)\}'$, and $\beta_l(s)$ is the unknown smooth function of s to be estimated nonparametrically. This modeling approach would allow complex interactions between S and $\mathbf{W}$. Here, we use the additional subscript m in $\widehat{\mu}_{0m}^{(p)}(\cdot, \cdot)$ to emphasize the fact that this estimator of $\mu_0(\cdot, \cdot)$ will now be fully or partially dependent on model assumptions, i.e., model-based. Certainly, given this model dependence, robustness (or lack thereof) to model misspecification would need to be carefully considered when using this approach in practice.

4 | SIMULATION STUDY

4.1 | Simulation Goals and Setup

The two main goals of our simulation study were: 1) to examine the finite sample properties of our estimation procedure for $\widetilde{\Delta}_H$ in terms of bias, accuracy of our variance calculation, and coverage of constructed confidence intervals, and 2) to compare testing results based on the three different testing quantities: $\widehat{\Delta}$ (using the primary outcome, gold standard) vs. $\widehat{\Delta}_P$ (using the surrogate marker, ignoring heterogeneity) vs. $\widehat{\Delta}_H$ (using the surrogate marker, accounting for heterogeneity). For the testing results, we focus on the point estimates themselves, the resulting effect sizes (point estimate/standard error estimate), and power. Importantly, when there is heterogeneity, we do not necessarily aim to demonstrate improved power with our proposed approach but rather, to demonstrate settings where the testing procedure using $\widehat{\Delta}_P$ (using the surrogate marker, ignoring heterogeneity) can be incorrect.

To achieve these goals, we examined eight simulation settings. For all settings, results were summarized over 500 replications; we examined all settings with $(n_1^p, n_0^p) = (1000, 800)$ (sample sizes in prior study) and $(n_1, n_0) = (300, 300)$ (sample sizes in current study). All simulation settings were also repeated with $(n_1^p, n_0^p) = (300, 300)$ (sample sizes in prior study) and $(n_1, n_0) = (300, 300)$; results were similar and are not shown here. In setting 1, we generated data such that there was heterogeneity in the utility of the surrogate with respect to a baseline covariate *and* the distribution of this baseline covariate was different in the current study compared to the prior study. Specifically, in the prior study, which is fixed

in all simulations, $W_{1i}^p \sim U(0, 10)$, $W_{0i}^p \sim U(0, 10)$, $S_{1i}^p \sim gamma(shape = 2.78, scale = 2.78)$, and $S_{0i}^p \sim gamma(shape = 2.5, scale = 2.5)$. We then generate the outcomes from:

$$Y_{1i}^p = I(W_{1i}^p < 5)(3.5 + 5S_{1i}^p) + I(W_{1i}^p \geq 5)(16S_{1i}^p) + N(0, 16),$$

$$Y_{0i}^p = I(W_{0i}^p < 5)(3.2 + 4S_{0i}^p) + I(W_{0i}^p \geq 5)(15.95S_{0i}^p) + N(0, 16).$$

where throughout $N(a, b)$ indicates a normal distribution with mean a and variance b. The motivation behind this setup was (a) to generate a surrogate marker where higher values are desirable and the surrogate level tends to be higher in the treated group, and (b) to generate an outcome where the surrogate marker is positively associated with the outcome but this association is stronger in magnitude in the treated group, reflecting residual treatment effect beyond the surrogate marker. In addition, to induce heterogeneity, we generate data such that the treatment effect on the primary outcome and the association between primary outcome and surrogate marker depend on whether the covariate is less than or greater than 5. With this setup, there was a statistically significant heterogeneity in surrogacy based on the test for heterogeneity proposed by Parast et al. (2021); the estimated proportion of treatment effect explained by the surrogate marker was 0.52 for $W_{gi}^p < 5$ and 0.95 for $W_{gi}^p \geq 5$, $g \in \{0, 1\}$. In this setting, the $(S_{gi}, Y_{gi}) \mid W_{gi}$ in the current study was generated the same as in the prior study, but W_{1i} and W_{0i} were generated from a $U(0,4)$, which is different from the prior study. Note that for all patients in the current study, the surrogate strength is not very strong and thus, we would expect that using the surrogate but ignoring heterogeneity will lead to an overestimation of the treatment effect. While the variability of the primary outcome, Y_{gi}, is large in both treatment groups, the size of the treatment effect is large as well. For example, in this setting, our results will show that the average estimated treatment effect on the outcome in the current study is 14.10, and the empirical power of testing the treatment effect is 100% using the primary outcome only.

In setting 2, W_{gi}^p and $Y_{gi}^p \mid S_{gi}^p, W_{gi}^p$ in the prior study were generated exactly the same as in setting 1, but $S_{1i}^p \sim gamma(shape = 2.66, scale = 2.66)$ and $S_{0i}^p \sim gamma(shape = 2.5, scale = 2.5)$. The motivation behind this change in the distributions for the surrogate marker is that we aimed to make the treatment effect on both the primary outcome and surrogate marker smaller than in setting 1, in order to explore how the various tests performed when less power would be expected. As in setting 1, there was significant heterogeneity in surrogacy with the estimated proportion of treatment effect explained by the surrogate being 0.39 for $W_{gi}^p < 5$ and 0.90 for $W_{gi}^p \geq 5$. The current study was generated the same as the prior study except that W_{1i} and W_{0i} were generated from a $U(6, 10)$ distribution. In contrast to setting 1, for all patients in the current study, the surrogate is strong and thus, we would expect that using the surrogate but ignoring heterogeneity will lead to an underestimation of the treatment effect. With respect to the size of the treatment effect and empirical power in this setting, our results will show that the average treatment effect on the outcome in the current study is 13.34 , and the empirical power of testing the treatment effect is 69% using the primary outcome only.

In setting 3, (W_{gi}, S_{gi}) in the prior study were generated as in setting 2, but $Y_{1i}^p = I(W_{1i}^p < 5)(3.5 + 5 \times 7) + I(W_{1i}^p \geq 5)(16S_{1i}^p) + N(0, 16)$ and $Y_{0i}^p = I(W_{0i}^p < 5)(3.2 + 4 \times 6.25) + I(W_{0i}^p \geq 5)(15.95S_{0i}^p) + N(0, 16)$. The motivation behind this change in the distributions for Y was to explicitly make the surrogate useless among those with $W_{gi}^p < 5$ i.e., a more extreme version of setting 2. As expected, there was significant surrogacy heterogeneity with the treatment effect on the surrogate marker not explaining any of the treatment effect on the primary outcome among patients with $W_{gi}^p < 5$, and explaining the majority of the treatment effect on the primary outcome among patients with $W_{gi}^p \geq 5$ (proportion explained ≈ 0.92). Similar to setting 2, the current study was generated the same as the prior study except that W_{1i} and W_{0i} were generated from a $U(6, 10)$ distribution and thus, we expect a potentially larger gain in power using our proposed approach (though again, this is not our primary goal). With respect to the size of the treatment effect and empirical power in this setting, our results will show that the average treatment effect on the primary outcome in the current study is 13.34 , and the empirical power of testing the treatment effect is 69% using the primary outcome only, parallel to setting 2.

In setting 4, the prior study was generated exactly the same as in setting 1, and the current study was generated exactly the same as the prior study, i.e., W_{1i} and W_{0i} were generated from a $U(0, 10)$ distribution. Here, even though there is heterogeneity as described above for setting 1, since the covariate distribution is the same in prior and current studies, we expect the tests ignoring vs. accounting for heterogeneity to produce similar results. With respect to the size of the treatment effect and empirical power in this setting, our results will show that the average treatment effect on the primary outcome in the current study is 19.12 , and the empirical power of testing the treatment effect is 96% using the primary outcome only.

In setting 5, data were generated such that there is no heterogeneity. Specifically, in the prior study, $W_{1i}^p \sim U(0, 10), W_{0i}^p \sim U(0, 10), S_{1i}^p \sim gamma(shape = 2.78, scale = 2.78)$, $S_{0i}^p \sim gamma(shape = 2.5, scale = 2.5), Y_{1i}^p = 3.5 + 5S_{1i}^p + N(0, 1)$, and $Y_{0i}^p = 3.2 + 4S_{0i}^p + N(0, 1)$, independent of the baseline covariate. The proportion of the treatment effect explained by the surrogate in the prior study was 0.47, which is homogeneous in the study population. Data from the current study was distributed the same as for the prior study. The purpose of this setting was to examine how the tests perform when there is no heterogeneity and no difference in distribution from the prior study to the current study. With respect to the size of the treatment effect and empirical power in this setting, our results will show that the average treatment effect on the outcome in the current study is 13.90 , and the empirical power of testing the treatment effect is 100% using the primary outcome only.

In setting 6, data are generated similar to setting 1 but with lower variability in the primary outcome resulting in a much larger effect size. In the prior study, $W_{1i}^p \sim U(0, 10), W_{0i}^p \sim U(0, 10), S_{1i}^p \sim gamma(shape = 3, scale = 3)$, $S_{0i}^p \sim gamma(shape = 2.1, scale = 2.2)$. For $W_{1i}^p < 5$ and $W_{0i}^p < 5, Y_{1i}^p = 3.5 + 5S_{1i}^p + N(0, 1)$, and $Y_{0i}^p = 1 + 3S_{0i}^p + N(0, 1)$, respectively. For $W_{1i}^p \geq 5$ and $W_{0i}^p \geq 5, Y_{1i}^p = 16S_{1i}^p + N(0, 1)$ and $Y_{0i}^p = 15.8S_{0i}^p + N(0, 1)$, respectively. There was a substantial heterogeneity in the utility of the surrogate with the proportion of treatment effect explained by the surrogate being 0.67 for $W_{gi}^p < 5$ and 0.98 for $W_{gi}^p \geq 5$. In the current study, the S and Y were generated the same

as in the prior study, but W_{1i} and W_{0i} were generated from a $U(0,4)$ distribution. As in setting 1, since the surrogate strength is not very strong in the current study, we would expect that using the surrogate but ignoring heterogeneity will lead to an overestimation of the treatment effect. With respect to the size of the treatment effect and empirical power in this setting, our results will show that the average treatment effect on the outcome in the current study is 33.70 , and the empirical power of testing the treatment effect is 100% using the primary outcome only.

Settings 7 and 8 reflect a null treatment effect setting and we include them so that we may examine the empirical Type 1 error rate. In both settings, data from the prior study are generated as $W_{gi}^p \sim U(0, 10)$, $S_{gi}^p \sim gamma(shape = 2.5, scale = 2.5)$, and $Y_{gi}^p = 3.2 + 4S_{gi}^p + N(0, 16)$ for $g = 0, 1$. That is, there is neither treatment effect on the surrogate marker nor the treatment effect on the primary outcome, and S_{gi} and Y_{gi} are positively associated. In setting 7, data in the current study are generated exactly as the prior study. In setting 8, data in the current study are generated such that $(S_{gi}, Y_{gi}) \mid W_{gi}$ are generated the same as the prior study, but $W_{gi} \sim U(0,4), g \in \{0, 1\}$, i.e., the distribution of the baseline covariate is different in the current study. The purpose of setting 8 is to specifically examine estimation and testing when there is no treatment effect and no heterogeneity, but the current study does have a different patient population compared to the prior study. In both settings, the true treatment effect on the primary outcome is 0 and the empirical Type 1 error of the test using the primary outcome is 0.06. In both settings, there is no empirical evidence that S is an "informative" surrogate marker, and no empirical evidence of heterogeneity in surrogacy, as expected.

With respect to our bandwidth selection, we let $h_0 = 1.06 \times \min(\sigma_{W_0}, IQR_0/1.34)n_0^{-2/5}$ and $h_1 = 1.06 \times \min(\sigma_{W_1}, IQR_0/1.34)n_1^{-2/5}$ where σ_{W_g} and IQR_g were the empirical standard deviation and inter-quartile range of W_g, and $h_2 = 2 \times 1.06 \times \min(\sigma_{S_0^p}, IQR_1/1.34)n_{0^p}^{-2/5}$ and $h_3 = 2 \times 1.06 \times \min(\sigma_{W_0^p}, IQR_2/1.34)n_{0^p}^{-2/5}$ where $\sigma_{S_0^p}$ and IQR_1 were the empirical standard deviation and inter-quartile range of S_{0^p}, respectively, and $\sigma_{W_{0^p}}$ and IQR_2 were the empirical standard deviation and inter-quartile range of W_{0^p}, and $h_4 = 1.06 \times \min(\sigma_{S_0^p}, IQR_1/1.34)n_{0^p}^{-0.31}$.[18,7]

4.2 | Simulation Results

Table 1 shows estimation results for $\widehat{\Delta}_H$ for all settings, using our proposed estimating procedure. We examine bias in coverage with respect to both $\widetilde{\Delta}_H$ (fixed prior study) and Δ_H. These results demonstrate good performance with minimal bias, average standard error estimates that are close to the empirical standard error, and coverage of the confidence intervals close to the nominal value of 95%.

Table 2 shows results from testing using $\widehat{\Delta}$, $\widehat{\Delta}_P$, and $\widehat{\Delta}_H$. In setting 1 where there is heterogeneity and the distribution of W in the current study is different from the prior study, results show that $\widehat{\Delta}_P$ overestimates the treatment effect and thus, does not retain the lower boundedness property. In contrast, our approach using $\widehat{\Delta}_H$ does not overestimate the treatment effect. The power using $\widehat{\Delta}_H$ is smaller than that using $\widehat{\Delta}$, but this is expected

since the data generation in this setting is such that the population in the current study is composed largely of individuals where the surrogate marker is not very strong. In setting 2 where there is again heterogeneity and the distribution of W in the current study is different from the prior study, results show that both $\widehat{\Delta}_P$ and $\widehat{\Delta}_H$ are less than $\widehat{\Delta}$, but $\widehat{\Delta}_H$ is much closer to $\widehat{\Delta}$ and has power equivalent to that using $\widehat{\Delta}$. This, again, is what was expected since the data generation in this setting is such that the population in the current study is composed largely of individuals where the surrogate marker is strong. In setting 3, which is similar to setting 2 but we have made the data more extreme with the surrogate being useless for those with $W < 5$, results show a larger departure in $\widehat{\Delta}_P$ from $\widehat{\Delta}$, and a larger decrease in power for $\widehat{\Delta}_P$ compared to $\widehat{\Delta}_H$. In setting 4 where there is heterogeneity but the distribution of W in both the prior study and the current study is the same, we see similar point estimates for $\widehat{\Delta}_P$ and $\widehat{\Delta}_H$ but a slightly higher standard error and lower power for $\widehat{\Delta}_H$. This indicates that in some settings, we may pay a price in terms of power and efficiency when we use the approach that accounts for heterogeneity when it is not necessary. In setting 5, where there is no heterogeneity, we see similar performance for $\widehat{\Delta}_P$ and $\widehat{\Delta}_H$. In setting 6, where we have a very large treatment effect on the primary outcome, there is heterogeneity and the distribution of W in the current study is different from the prior study, results show that, as expected, $\widehat{\Delta}_P$ overestimates the treatment effect and does not retain the lower boundedness property, as in setting 1. In settings 7 and 8, where there is no treatment effect, results show that all three testing procedures perform well with an estimated treatment effect close to zero and Type 1 error rate close to 0.05. We additionally examined the efficiency gain comparing our proposed estimator to the simple estimator in (4); indeed, we did observe efficiency gains using our proposed estimator, quantified by the ratio of the estimated standard error using our proposed estimate to that using the simple estimate, that ranged from 0.79-0.98 across settings.

In summary, results from this simulation study show 1) good finite sample performance of our estimation and inference procedures for Δ_H, 2) a potential slight loss in power when using the proposed $\widehat{\Delta}_H$ compared to $\widehat{\Delta}_P$ when accounting for heterogeneity is not needed, and 3) a potential for inaccurate conclusions and/or loss in power when $\widehat{\Delta}_P$ is used instead of the proposed $\hat{\Delta}_H$ when accounting for heterogeneity is needed.

5 | APPLICATION

We apply our proposed approach to test for a treatment effect based on a heterogeneous surrogate using data from two distinct AIDS clinical trials, the AIDS Clinical Trials Group (ACTG) 320 Study and the ACTG 193A Study.[19,20] These data are publicly available upon request from the AIDS Clinical Trial Group[21]. We consider the ACTG 320 Study as our prior study and the ACTG 193A Study as our current study. The ACTG 320 study was conducted among HIV-infected patients with a CD4 cell count of 200 or less per cubic millimeter and was a randomized, double-blind trial that compared a two-drug regimen (two nucleoside reverse transcriptase inhibitors [NRTI]) with a three-drug regimen (two NRTIs plus indinavir). There were a total of 830 participants, with 412 in the two-drug regimen group and 418 in the three-drug regimen group. The ACTG 193A study was a randomized,

double-blind trial conducted among HIV-infected patients with a CD4 cell count of 50 or less per cubic millimeter. We focus on the comparison of a two-drug regimen (NRTIs) with a three-drug regimen (two NRTIs plus nevirapine). There were a total of 657 participants, with 327 in the two-drug regimen group and 330 in the three-drug regimen group. Our primary outcome Y is the change in plasma HIV-1 RNA from baseline to 24 weeks; our surrogate marker S is change in CD4 cell count from baseline to 24 weeks, as CD4 is relatively less expensive to measure compared to RNA.[22] Both Y and S are available in ACTG 320 while only S is available in the publicly available data of ACTG 193A. Previous work has demonstrated significant heterogeneity in the utility of S with respect to W, baseline CD4 count, with the surrogate strength being stronger among those with a lower baseline CD4 count and weaker among those with a higher baseline CD4 count[12] as shown in Figure 1. We aim to use our proposed method to test for a treatment effect on RNA using CD4 count as a surrogate marker, accounting for the known heterogeneity in the utility of the surrogate which was demonstrated in the prior study.

In Figure 2 we show the distribution of the baseline covariate, baseline CD4, in the prior study compared to the current study. Clearly, the current study is composed of a different participant population with lower CD4 counts due to the study eligibility criteria. In Figure 1, we also see that the surrogate is strongest in this subgroup. Using our proposed approach, we obtain a treatment effect estimate of $\widehat{\Delta}_H = -0.10$ (standard error [SE] = 0.03) with a p-value < 0.001. Note that since *lower* plasma HIV-1 RNA is better, a negative change in RNA indicates a beneficial treatment effect for the three-drug regimen. Using the approach that does not account for heterogeneity, we obtain a treatment effect estimate closer to the null, but still significant: $\widehat{\Delta}_P = -0.07 (SE = 0.02), p < 0.001$. That is, while the overall conclusion regarding the treatment effect based on the surrogate would be significant using either test, our proposed test provides a treatment effect point estimate that is larger in magnitude. This is expected since the surrogate strength is greater in this subgroup that makes up the current study, and our proposed approach takes advantage of that information.

6 | DISCUSSION

For settings where it is known that the strength of a surrogate marker varies by a certain baseline characteristic, we have proposed an approach and estimation procedures to appropriately test for a treatment effect using only the surrogate marker, accounting for this known heterogeneity. We demonstrated good finite sample performance of our estimation procedure and showed that our proposed testing procedure can outperform an approach that does not account for heterogeneity. An R package implementing the methods proposed here, named `hettest`, is available at https://github.com/laylaparast/hettest.

While we largely focus, specifically in the numerical studies, on settings where the distribution of W is different in the current study as compared to the prior study, it is still possible for a test based on $\widehat{\Delta}_P$, i.e., ignoring heterogeneity, to provide inaccurate results about the treatment effect when there is heterogeneity in the utility of the surrogate and the W is distributed the *same* in the two studies; we provide an example in Appendix E.

In the presence of heterogeneity, both the treatment effect and the utility of the surrogate marker may depend on W. While we focus exclusively on the average treatment effect in this paper, it may be of interest to test for a treatment effect based on alternative summaries that account for such heterogeneity. For example, one may define $\Delta_w = E\left(Y^{(1)} \mid W^{(1)} = w\right) - E\left(Y^{(0)} \mid W^{(0)} = w\right)$ and the subgroup specific earlier treatment effect $\Delta_H(w) = \int \mu_0^p(s, w) dF^{(1)}(s \mid w) - \int \mu_0^p(s, w) dF^{(0)}(s \mid w)$. Then we may test for a treatment effect based on S by examining a functional of $\Delta_H(w)$ such as $\sup_w \Delta_H(w)$ or $\int \Delta_H(w) dw$, the area under the curve produced by $\Delta_H(w)$. Such alternative summaries of the treatment effect across a baseline covariate, W, are not unique to the surrogate marker setting as they have been extensively discussed in the general heterogeneous treatment effect literature.[23,24] However, these alternative summaries may also prove useful in the heterogeneous surrogate setting and may offer new insights over simply looking at the average treatment effect.

Importantly, we require Assumptions (C1) – (C4) and in practice, they may be violated. Specifically, if the existing heterogeneity is such that the surrogate is not strong or, worse, the treatment effect on the surrogate marker and primary endpoint may be in different directions for some w, the surrogate should *not* be used as a replacement of the primary outcome for all individuals in a future study. Instead, one may consider using the surrogate as a replacement only among those with a w where assumptions (C1) – (C4) hold. To achieve this, one could consider first identifying a region of interest where the surrogacy is sufficiently strong e.g., Ω_w such that the conditional average treatment effect on the primary endpoint $\Delta(w) \geq \delta_0 > 0$ and the proportion explained by the surrogate for $W = w$, $R_S(w) = \Delta_H(w) / \Delta(w)$, is between 0.50 and 1.0, and then apply the proposed testing procedure that replaces Y with S for testing the average treatment effect in the subpopulation Ω_w. If one is interested in studying the average treatment effect in the entire study population, one may combine the proposed test statistic with a new but simple test statistic measuring the strength of the treatment effect based on actual primary endpoints Y for patients in the complement of Ω. Such a hybrid approach has the potential to reduce costs if S is less costly to measure than Y and/or reduce the follow-up time needed for those in Ω_w if S is measured earlier than Y. Though not exactly within this context, previous work has explored the potential for auxiliary information (including but not limited to surrogate markers) to improve efficiency when testing for a treatment or intervention effect.[25,26] While this is beyond the scope of this paper, further work on this topic within the framework of a heterogeneous surrogate is warranted.

Our proposed approach has some limitations. First, if the current study includes participants with w values outside the observed distribution in the prior study, our approach will not be able to obtain $\hat{\mu}_0^p(s, w)$ for that w without extrapolation. In such a case, when there is observed heterogeneity in the prior study, use of the surrogate marker to test for a treatment effect in the current study should likely be limited to those with w contained in the prior study. Second, given our use of kernel smoothing, we require a relatively large sample size. Robust nonparametric methods for surrogate markers are lacking in general for small sample size settings; future work in this area would be needed. Lastly, we require several assumptions, outlined in Section 2.2, which are generally untestable though they

may be empirically explored using the observed data. These assumptions are needed for identifiability, to ensure our lower-boundedness property of Δ_H (i.e., $\Delta_H \leq \Delta$), and to guard against the surrogate paradox which occurs when the surrogate and outcome are positively associated, the treatment has a positive effect on the surrogate, but the treatment in fact has a negative effect on the outcome.[15] The surrogate paradox is especially of concern here as our primary goal is to make a conclusion about the treatment effect on the primary outcome based on information about the surrogate marker. While these assumptions are strong, they are more likely to hold than the parallel assumptions required for Δ_P [7] to be valid due to the additional conditioning on W. Further work on methods that allow for more relaxed assumptions and/or that allow one to assess sensitivity to violations of these assumptions would be useful.[27]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

APPENDIX

APPENDIX A

Discrete Example

Let Y denote the primary outcome and S denote the surrogate marker. We use potential outcomes notation where each person has a potential $\left\{Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\right\}$ where $Y^{(g)}$ and $S^{(g)}$ are the outcome and surrogate when the patient receives treatment g. Our main quantity of interest is the treatment effect on the primary outcome quantified as $\Delta \equiv E\left(Y^{(1)} - Y^{(0)}\right) = E\left(Y^{(1)}\right) - E\left(Y^{(0)}\right)$. The earlier treatment effect incorporating S information is defined in the main text as

$$\Delta_P = \int \mu_0^p(s) dF^{(1)}(s) - \int \mu_0^p(s) dF^{(0)}(s) \tag{1}$$

where $\mu_0^p(s) \equiv E\left(Y^{(0p)} = y \mid S^{(0p)} = s\right)$. In this example, we will have heterogeneity in the utility of the surrogate with respect to gender. Consider our prior study, which we refer to as Study A in this example, and is shown in Figure 1. The Study A sample is 50% female and 50% male. For all individuals, $\left(S^{(1)}, S^{(0)}\right)$ are independent of gender, and $\left\{E\left(S^{(1)}\right), E\left(S^{(0)}\right)\right\} = (10,5)$. For females, $E\left(Y^{(1)} \mid S^{(1)} = s\right) = 3 + 5s$ and $E\left(Y^{(0)} \mid S^{(0)} = s\right) = 1 + 3S$. It can be shown that for females, $\Delta = 53 - 16 = 37$ and $\Delta_P = 15$. The proportion of the treatment effect on the primary outcome that is explained by the surrogate among females is thus 15/37=41%, which would not be considered as a strong

surrogacy. For males, $E\left(Y^{(1)} \mid S^{(1)} = s\right) = 15s$ and $\left(Y^{(0)} \mid S^{(0)} = s\right) = 14.8S$. It can be shown that for males, $(\Delta, \Delta_P) = (76, 74)$ and the proportion explained by the surrogate marker is 97% among males, representing strong surrogacy.

To calculate Δ_P for a future study, let's consider the conditional mean that is central to this calculation, $\mu_0^p(s) = E\left(Y^{(0p)} = y \mid S^{(0p)} = s\right)$ where the superscript p indicates that this is referring to the prior study, i.e., study A. In this example, this would be $\mu_0^p(s) = 0.5 \times (1 + 3s) + 0.5 \times 14.8s = 8.9s + 0.5$. Now assume our current study is Study B shown in Figure 1 which is 95% female and 5% male. Importantly, the joint distributions of $\left(Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\right)$ in males and females remain as described above; the only difference is the distribution of gender. The treatment effect, Δ in this new study is 0.95×37+0.05×76=38.95. If one were to calculate Δ_P not accounting for this known heterogeneity in the utility of the surrogate, the quantity obtained would be $\Delta_P = 8.9 \times 10 + 0.5 - 8.9 \times 5 - 0.5 = 44.5$, recalling that $E\left(S^{(1)}\right) = 10$ and $E\left(S^{(0)}\right) = 5$ for all individuals in both studies. However, using our proposed approach which does account for heterogeneity, we use Δ_H as the earlier treatment effect, defined in the main text as:

$$\Delta_H = \int \mu_0^p(s, w) dF^{(1)}(s, w) - \int \mu_0^p(s, w) dF^{(0)}(s, w).$$

Thus, $\Delta_H = 95\% \times (1 + 3 \times 10) + 5\% \times (14.8 \times 10) - 95\% \times (1 + 3 \times 5) - 5\% \times (14.8 \times 5) = 17.95$. Therefore $\Delta_H < \Delta < \Delta_P$ and Δ_P no longer retains the property of providing a lower bound on the treatment effect on Y.

Now we consider a study, labeled Study C in Figure 1, which is 95% males and 5% females. Using similar calculations, we can show that $\Delta = 74.05$, $\Delta_P = 44.05$ and $\Delta_H = 71.05$. Thus, in this case, Δ_H will provide better lower bound for Δ and the test based on Δ_H is expected to be more powerful than that based on Δ_P. The discrete case, as illustrated in this example, is relatively straightforward in terms of how to go about calculating the needed quantities separately by group and appropriately accounting for the different distribution in the new study. The continuous baseline covariate case, however, is more complex, and our Appendix C presents an example such that even if the prior and current studies have the same distribution for covariates, Δ_P may still fail to be a valid lower bound for Δ.

APPENDIX B

As noted in this text, Assumptions $(C1) - (C3)$ together guarantee that $E\left(Y^{(1)} \mid W = w\right) \geq E\left(Y^{(0)} \mid W = w\right)$, for all w in the support of W. This result is due to the derivation:

$$
\begin{aligned}
\Delta\,(w) &= E\!\left(Y^{(1)} \mid W = w\right) - E\!\left(Y^{(0)} \mid W = w\right)\\
&= \int_s E\!\left(Y^{(1)} \mid S^{(1)} = s, W = w\right)dF^{(1)}(s \mid w) - \int_s E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)dF^{(0)}(s \mid w)\\
&\geq \int_s E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)dF_1(s \mid w) - \int_s E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)dF^{(0)}(s \mid w)\\
&= \int_s E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)d\!\left\{F^{(1)}(s \mid w) - F^{(0)}(s \mid w)\right\}\\
&= \int_s \left\{F^{(0)}(s \mid w) - F^{(1)}(s \mid w)\right\}\frac{\partial E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right)}{\partial s}ds \geq 0,
\end{aligned}
$$

where $F^{(g)}\!\left(s \mid w\right) = P\!\left(S^{(g)} \leq s \mid W = w\right), g = 0, 1$. That is, while treatment effect heterogeneity is allowed, the directions of the conditional average treatment effect among subgroups of patients with $W = w$ need to be consistent. One important implication is that under the null H_0: $\Delta\, = E\{\,\Delta\,(W)\} = 0$, i.e., no average treatment effect, the conditional average treatment effect $\Delta\,(w) = 0$ for all w as well. Furthermore, from the derivation, it is clear that $\Delta\,(w) = 0$ if and only if both

1. $F^{(1)}\!\left(s \mid w\right) = F^{(0)}\!\left(s \mid w\right)$, i.e., $P\!\left(S^{(1)} > s \mid W = w\right) = P\!\left(S^{(0)} > s \mid W = w\right)$ and

2. $E\!\left(Y^{(1)} \mid S^{(1)} = s, W = w\right) = E\!\left(Y^{(0)} \mid S^{(0)} = s, W = w\right).$

Specifically, $\Delta\,(w) = 0$ implies that there is no treatment effect on the distribution of the surrogate marker in the subgroup of patients with $W = w$. In summary, under Assumptions $(C1)$-$(C3)$

$$
\Delta\, = 0 \Rightarrow \Delta\,(w) = 0 \Rightarrow S^{(1)}\big|W = w \sim S^{(0)}\big|W = w\,.
$$

This relationship allows us to test the common null H_0: $\Delta\, = 0$ via testing a seemingly more restrictive null that $S^{(1)} \mid W = w \sim S^{(0)} \mid W = w$, for all w in the support of W.

For $(C2)$ and $(C3)$, if the primary outcome or surrogate are such that lower values are "better", one can simply define the outcome/surrogate as $-X$ where X is the initial value.

Assumptions $(C5) - (C6)$ are not required for the validity of the testing procedure proposed in the next section in that the p-value under the null follows a uniform distribution even without them, but it allows us to estimate a lower bound of the average treatment effect, Δ, and construct the corresponding test statistic.

Under the following additional assumptions:

(C7) $Y^{(1)} \perp S^{(0)} \mid S^{(1)}, W$ and $Y^{(0)} \perp S^{(1)} \mid S^{(0)}, W$;

(C8) $Y^{(1p)} \perp S^{(0p)} \mid S^{(1p)}, W^p$ and $Y^{(0p)} \perp S^{(1p)} \mid S^{(0p)}, W^p,$

the treatment effect on the surrogate marker defined in Section 2.3 and on the primary outcome can be interpreted within a causal framework: the proposed test statistic is an estimate of the portion of the treatment effect on the primary outcome attributable to the treatment effect on the surrogate marker. Otherwise, the proposed treatment effect on the surrogate marker can always serve as a lower bound for the average treatment effect on Y and can be used in practice without assuming them.

To summarize, Assumptions $(C1) – (C4)$ are needed for the validity of the proposed testing procedure, Assumptions $(C5) – (C6)$ allow us to interpret the test statistic based on he surrogate marker and baseline covariate only as a "conservative" estimator (or a lower bound) of the average treatment effect on the primary outcome, and causal interpretation of the lower is possible under additional assumptions $(C7) – (C8)$.

APPENDIX C

To estimate Δ using the primary outcome (gold standard) we use $\widehat{\Delta} = n_1^{-1} \sum_{i=1}^{n_1} Y_{1i} - n_0^{-1} \sum_{i=1}^{n_0} Y_{0i}$ and conduct a t-test to test H_0: $\Delta = 0$.

To estimate $\widetilde{\Delta}_P$, we use the nonparametric estimation approach of[7] by estimating $\mu_0^p(s)$ as

$$\widehat{\mu}_0^p(s) = \frac{\sum_{i=1}^{n_0^p} K_{h_4}(S_{0i}^p - s) Y_{0i}^p}{\sum_{i=1}^{n_0^p} K_{h_4}(S_{0i}^p - s)},$$

and then estimate $\widehat{\Delta}_P$ as

$$\widehat{\Delta}_P = n_1^{-1} \sum_{i=1}^{n_1} \widehat{\mu}_0^p(S_{1i}) - n_0^{-1} \sum_{i=1}^{n_0} \widehat{\mu}_0^p(S_{0i}).$$

Note that this estimate only uses S data from the current study (no Y data from the current study) and S, Y data from the previous study in group $Z = 0$ only. To obtain an estimate for the standard error of $\widehat{\Delta}_P$, σ_P, we simply take the empirical standard deviation of the transformed surrogate i.e., let $\widetilde{Y}_{gi} = \widehat{\mu}_0^p(S_{gi})$, and then $\widehat{\sigma}_P = \widehat{var}(\widetilde{Y}_{1i})/n_1 + \widehat{var}(\widetilde{Y}_{0i})/n_0$ where $\widehat{var}$ indicates the empirical variance. This alternative testing procedure would then use the test statistic $Z_P = \widehat{\Delta}_P / \widehat{\sigma}_P$ and reject the null hypothesis when $|Z_P| > \Phi^{-1}(1 - \alpha/2)$.

Importantly, one may also consider simply using the surrogate markers measured in the current study and define $\Delta_M = E(S^{(1)}) - E(S^{(0)})$ and conduct a t-test of H_{0M}: $\Delta_M = 0$. The disadvantage of this approach is that there is no way to relate Δ_M and Δ i.e., the estimate of Δ_M does not give any helpful information about the magnitude of Δ. In addition, this approach does not take advantage of information from the previous study nor does it account for heterogeneity in the utility of the surrogate marker. For these reasons, we do not compare our approach to this test.

APPENDIX D

Our proposed estimator for $\widetilde{\Delta}_H$ is

$$\widehat{\Delta}_H = \frac{1}{n}\left\{\sum_{i=1}^{n_0}[\widehat{m}_1(W_{0i};\widehat{\mu}_0^p) - \widehat{m}_0(W_{0i};\widehat{\mu}_0^p)] + \sum_{i=1}^{n_1}[\widehat{m}_1(W_{1i};\widehat{\mu}_0^p) - \widehat{m}_0(W_{1i};\widehat{\mu}_0^p)]\right\}.$$

Let $\widetilde{\mu}_g = E\left\{\widehat{\mu}_0^p\left(S^{(g)}, W\right) \mid \widehat{\mu}_0^p\right\}, g = 0, 1$. It is obvious that $\widetilde{\Delta}_H = \widetilde{\mu}_1 - \widetilde{\mu}_0$. Also, let $m_g(w;\widehat{\mu}_0^p) = E\left\{\widehat{\mu}_0^p\left(S^{(g)}, W\right) \mid W = w\right\}$.

In this section, we only consider the randomness in the current study, i.e., the probability measure is conditional on $\widehat{\mu}_0^p(\cdot, \cdot)$. Now consider the centered term

$$\frac{1}{n}\sum_{g=0}^{1}\sum_{j=1}^{n_g}\widehat{m}_1(W_{gj};\widehat{\mu}_0^p) - \widetilde{\mu}_1$$
$$= \frac{1}{n}\sum_{g=0}^{1}\sum_{j=1}^{n_g}\left[n_1^{-1}\sum_{i=1}^{n_1}\frac{K_h(W_{1i} - W_{gj})\widetilde{S}_{1i}}{\widehat{f}_1(W_{gj})}\right] - \widetilde{\mu}_1,$$

which is

$$\frac{1}{nn_1}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\frac{K_h(W_{1i} - W_{0j})\widetilde{S}_{1i}}{\widehat{f}_1(W_{0j})} + \frac{1}{n}\sum_{i=1}^{n_1}\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{K_h(W_{1i} - W_{1j})}{\widehat{f}_1(W_{1j})}\right]\widetilde{S}_{1i} - \widetilde{\mu}_1$$

$$= \frac{1}{nn_1}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\frac{K_h(W_{1i} - W_{0j})\widetilde{S}_{1i}}{\widehat{f}_1(W_{0j})} + \frac{1}{n}\sum_{i=1}^{n_1}\left[\frac{1}{n_1}\sum_{j=1}^{n_1}K_h(W_{1i} - W_{1j})\right]\frac{\widetilde{S}_{1i}}{\widehat{f}_1(W_{1i})} - \widetilde{\mu}_1 + O_p\left(h^2\right)$$

$$= \frac{n_0}{nn_1}\sum_{i=1}^{n_1}\frac{\widehat{f}_0(W_{1i})}{\widehat{f}_1(W_{1i})}\widetilde{S}_{1i} + \frac{1}{n}\sum_{i=1}^{n_1}\widetilde{S}_{1i} - \widetilde{\mu}_1 + O_p\left(h^2\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widetilde{S}_{1i} - \widetilde{\mu}_1\right) + \frac{n_0}{nn_1}\sum_{i=1}^{n_1}\frac{\widehat{f}_0(W_{1i}) - \widehat{f}_1(W_{1i})}{\widehat{f}_1(W_{1i})}\widetilde{S}_{1i} + O_p\left(h^2\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widetilde{S}_{1i} - \widetilde{\mu}_1\right) + \frac{n_0}{nn_1}\sum_{i=1}^{n_1}\left[\frac{1}{n_0}\sum_{j=1}^{n_0}K_h(W_{0j} - W_{1i}) - \frac{1}{n_1}\sum_{j=1}^{n_1}K_h(W_{1j} - W_{1i})\right]\frac{\widetilde{S}_{1i}}{f_1(W_{1i})} + O_p\left(h^2\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widetilde{S}_{1i} - \widetilde{\mu}_1\right) + \pi_0\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{m}_1(W_{0i};\widehat{\mu}_0^p) - \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{m}_1(W_{1i};\widehat{\mu}_0^p)\right] + O_p\left(h^2\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widetilde{S}_{1i} - \widetilde{\mu}_1\right) + \pi_0\left[\frac{1}{n_0}\sum_{i=1}^{n_0}m_1(W_{0i};\widehat{\mu}_0^p) - \frac{1}{n_1}\sum_{i=1}^{n_1}m_1(W_{1i};\widehat{\mu}_0^p)\right]$$
$$+ \pi_0\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\widehat{m}_1(W_{0i};\widehat{\mu}_0^p) - m_1(W_{0i};\widehat{\mu}_0^p)\right) - \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{m}_1(W_{1i};\widehat{\mu}_0^p) - m_1(W_{1i};\widehat{\mu}_0^p)\right)\right] + O_p\left(h^2\right)$$

where $\pi_g = n_g/n$ and $\widehat{f}_1(w)$ is the nonparametric estimator for the density function of W based on observations in treatment group 1. Now, consider the expansion

$$\widehat{m}_1(w;\widehat{\mu}_0^p) - m_1(w;\widehat{\mu}_0^p) = \frac{1}{n_1}\sum_{i=1}^{n_1}K_h(W_{1i} - w)\left\{\widetilde{S}_{1i} - m_1(W_{1i};\widehat{\mu}_0^p)\right\} + O_p\left(h^2 + \frac{\log(n_1)}{n_1 h}\right)$$

uniform in w. Therefore,

$$\frac{1}{n_0}\sum_{j=1}^{n_0}\left\{\widehat{m}_1(W_{0j};\widehat{\mu}_0^p) - m_1(W_{0j};\widehat{\mu}_0^p)\right\}$$

$$= \frac{1}{n_1 n_0}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1} K_h(W_{1i} - W_{0j})\left\{\widetilde{S}_{1i} - m_1(W_{1i};\widehat{\mu}_0^p)\right\} + O_p\left(h^2 + \frac{\log(n_1)}{n_1 h}\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_0}\widehat{f}_0(W_{1i})\left\{\widetilde{S}_{1i} - m_1(W_{1i};\widehat{\mu}_0^p)\right\} + O_p\left(h^2 + \frac{\log(n_1)}{n_1 h}\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_0} f_0(W_{1i})\left\{\widetilde{S}_{1i} - m_1(W_{1i};\widehat{\mu}_0^p)\right\} + O_p\left(h^2 + \frac{\log(n_1)}{n_1 h}\right) + o_p\left(\frac{1}{\sqrt{n_1}}\right)$$

Similarly,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{m}_1(W_{1i};\widehat{\mu}_0^p) - m_1(W_{1i};\widehat{\mu}_0^p)\right)$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_0} f_0(W_{1i})\left\{\widetilde{S}_{1i} - m_1(W_{1i};\widehat{\mu}_0^p)\right\} + O_p\left(h^2 + \frac{\log(n_1)}{n_1 h}\right) + o_p\left(\frac{1}{\sqrt{n_0}}\right),$$

and

$$\sqrt{n}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\widehat{m}_1(W_{0i};\widehat{\mu}_0^p) - m_1(W_{0i};\widehat{\mu}_0^p)\right) - \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{m}_1(W_{1i};\widehat{\mu}_0^p) - m_1(W_{1i};\widehat{\mu}_0^p)\right)\right] \tag{2}$$

$$= O_p\left(\sqrt{n_1}h^2 + \frac{\log(n_1)}{\sqrt{n_1}h}\right) + o_p(1). \tag{3}$$

Therefore, when $h = O\left(n_1^{-\delta}\right), \delta \in \left(1/4, 1/2\right)$, the right hand side of (3) becomes $o_p(1)$, and thus

$$\frac{1}{\sqrt{n}}\sum_{g=0}^{1}\sum_{j=1}^{n_g}\widehat{m}_1(W_{gj};\widehat{\mu}_0^p) - \widetilde{\mu}_1$$

$$= \frac{\sqrt{n}}{n_1}\sum_{i=1}^{n_1}\left(\widetilde{S}_{1i} - \widetilde{\mu}_1\right) + \pi_0\left[\frac{\sqrt{n}}{n_0}\sum_{j=1}^{n_0} m_1(W_{0j};\widehat{\mu}_0^p) - \frac{\sqrt{n}}{n_1}\sum_{j=1}^{n_1} m_1(W_{1j};\widehat{\mu}_0^p)\right] + o_p(1).$$

Finally, we have

$$\sqrt{n}\left\{ \widehat{\Delta}_H - \widetilde{\Delta}_H \right\}$$

$$= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} (\widetilde{S}_{1i} - \widetilde{\mu}_1) + \pi_0 \left[\frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_0} m_1(W_{0i}; \widehat{\mu}_0^p) - \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} m_1(W_{1i}; \widehat{\mu}_0^p) \right] - \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_0} (\widetilde{S}_{0i} - \widetilde{\mu}_0)$$

$$+ \pi_1 \left[\frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} m_0(W_{1i}; \widehat{\mu}_0^p) - \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_0} m_0(W_{0i}; \widehat{\mu}_0^p) \right] + o_p(1)$$

$$= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \left(\widetilde{S}_{1i} - \pi_0 m_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1(\widetilde{\mu}_1 - \widetilde{\mu}_0) \right)$$

$$- \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_0} \left(\widetilde{S}_{0i} - \pi_0 m_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0(\widetilde{\mu}_1 - \widetilde{\mu}_0) \right) + o_p(1),$$

which converges weakly to a mean zero Gaussian distribution with a variance of

$$\frac{1}{\pi_1} E\left\{ \widetilde{S}_{1i} - \pi_0 m_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H \right\}^2 + \frac{1}{\pi_0} E\left\{ \widetilde{S}_{0i} - \pi_0 m_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widetilde{\Delta}_H \right\}^2 .$$

Therefore, the variance of $\widehat{\Delta}_H$ can be estimated as

$$\widehat{\sigma}_H^2 = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \left(\widetilde{S}_{1i} - \pi_0 \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widehat{\Delta}_H) \right)^2$$

$$+ \frac{1}{n_0^2} \sum_{i=1}^{n_0} \left(\widetilde{S}_{0i} - \pi_0 \widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widehat{\Delta}_H \right)^2$$

Next, we will derive the asymptotical distribution of $\sqrt{n}\left(\widehat{\Delta}_H^{AUG} - \widetilde{\Delta}_H \right)$. It is clear that

$$\sqrt{n}\left(\widehat{\Delta}_H^{AUG} - \widetilde{\Delta}_H \right)$$

$$= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{1i} - \pi_0 \widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H \right\}$$

$$- \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{0i} - \pi_0 \widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 \widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widehat{\Delta}_H \right\}$$

$$= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{1i} - \pi_0 m_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H \right\}$$

$$- \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{0i} - \pi_0 m_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widetilde{\Delta}_H \right\}$$

$$- \sqrt{n}\left[\frac{\pi_0}{n_0} \sum_{i=1}^{n_0} (\widehat{m}_1(W_{0i}; \widehat{\mu}_0^p) - m_1(W_{0i}; \widehat{\mu}_0^p)) - \frac{\pi_0}{n_1} \sum_{i=1}^{n_1} (\widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - m_1(W_{1i}; \widehat{\mu}_0^p)) \right]$$

$$- \sqrt{n}\left[\frac{\pi_1}{n_1} \sum_{i=1}^{n_1} (\widehat{m}_1(W_{1i}; \widehat{\mu}_0^p) - m_1(W_{1i}; \widehat{\mu}_0^p)) - \frac{\pi_1}{n_0} \sum_{i=1}^{n_0} (\widehat{m}_0(W_{0i}; \widehat{\mu}_0^p) - m_1(W_{0i}; \widehat{\mu}_0^p)) \right]$$

$$= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{1i} - \pi_0 m_1(W_{1i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \widehat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H \right\}$$

$$- \frac{\sqrt{n}}{n_0} \sum_{i=1}^{n_1} \left\{ \widetilde{S}_{0i} - \pi_0 m_1(W_{0i}; \widehat{\mu}_0^p) - \pi_1 m_0(W_{0i}; \widehat{\mu}_0^p) - \pi_0 \widetilde{\Delta}_H \right\} + o_p(1)$$

$$= \sqrt{n}\left(\widehat{\Delta}_H - \widetilde{\Delta}_H \right) + o_p(1).$$

Therefore, $\widehat{\Delta}_H^{AUG}$ and $\widehat{\Delta}_H$ are asymptotically equivalent. Furthermore, noting that

$$\tilde{S}_{1i} - \pi_0 m_1(W_{1i}; \hat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \hat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H$$
$$= \left\{ \tilde{S}_{1i} - m_1(W_{1i}; \hat{\mu}_0^p) \right\} + \pi_1 \left\{ m_1(W_{1i}; \hat{\mu}_0^p) - m_0(W_{1i}; \hat{\mu}_0^p) - \widetilde{\Delta}_H \right\}$$

and

$$\left[\left\{ \tilde{S}_{1i} - m_1(W_{1i}; \hat{\mu}_0^p) \right\} \left\{ m_1(W_{1i}; \hat{\mu}_0^p) - m_0(W_{1i}; \hat{\mu}_0^p) - \widetilde{\Delta}_H \right\} \mid W_{1i} \right] = 0,$$

we have

$$E\left[\tilde{S}_{1i} - \pi_0 m_1(W_{1i}; \hat{\mu}_0^p) - \pi_1 m_0(W_{1i}; \hat{\mu}_0^p) - \pi_1 \widetilde{\Delta}_H \right]^2$$
$$= E\left[\tilde{S}_{1i} - m_1(W_{1i}; \hat{\mu}_0^p) \right]^2 + \pi_1^2 E\left[m_1(W_{1i}; \hat{\mu}_0^p) - m_0(W_{1i}; \hat{\mu}_0^p) - \widetilde{\Delta}_H \right]^2.$$

Similarly,

$$E\left[\tilde{S}_{0i} - \pi_0 m_1(W_{0i}; \hat{\mu}_0^p) - \pi_1 m_0(W_{0i}; \hat{\mu}_0^p) - \pi_0 \widetilde{\Delta}_H \right]^2$$
$$= E\left[\tilde{S}_{0i} - m_0(W_{0i}; \hat{\mu}_0^p) \right]^2 + \pi_0^2 E\left[m_1(W_{0i}; \hat{\mu}_0^p) - m_0(W_{0i}; \hat{\mu}_0^p) - \widetilde{\Delta}_H \right]^2.$$

Therefore, the variance of $\widehat{\Delta}_H^{(AUG)}$ can also be consistently estimated by

$$\hat{\sigma}_{AUG}^2 = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \left[\hat{\mu}_0^{(p)}(S_{1i}, W_{1i}) - \hat{m}_1(W_{1i}; \hat{\mu}_0^p) \right]^2 + \frac{1}{n_0^2} \sum_{i=1}^{n_0} \left[\hat{\mu}_0^{(p)}(S_{0i}, W_{0i}) - \hat{m}_0(W_{0i}; \hat{\mu}_0^p) \right]^2$$
$$+ \frac{\pi_1^2}{n_1^2} \sum_{i=1}^{n_1} \left[\hat{m}_1(W_{1i}; \hat{\mu}_0^p) - \hat{m}_0(W_{1i}; \hat{\mu}_0^p) - \widehat{\Delta}_H \right]^2 + \frac{\pi_0^2}{n_0^2} \sum_{i=1}^{n_0} \left[\hat{m}_1(W_{0i}; \hat{\mu}_0^p) - \hat{m}_0(W_{0i}; \hat{\mu}_0^p) - \widehat{\Delta}_H \right]^2,$$

and $\widehat{\Delta}_{(AUG)} / \widehat{\Delta}_H = 1 + o_p(1)$.

APPENDIX E

Here, we provide an example where there is heterogeneity in the utility of the surrogate and the W is distributed the *same* in the prior study and current study, but Δ_P still fails to provide a lower bound for Δ. In both the prior study and the current study, we assume that $\log(W) \sim \epsilon_W$, $S^{(g)} = W \times \exp(\delta_0 g + \epsilon_S)$, and $Y^{(g)} = S^{(g)} W, g \in \{0, 1\}$, where δ_0 is a positive constant, and ϵ_W and ϵ_S are two independent standard normals. It is obvious that $\mu_0^p(s, w) = sw$ and

$$\Delta = \Delta_H = E\left(S^{(1)} W\right) - E\left(S^{(0)} W\right) = E\left\{ W E\left(S^{(1)} - S^{(0)} | W\right) \right\}$$
$$= E\{W(\exp(0.5 + \delta_0)W - \exp(0.5)W)\} = \exp\left(\frac{5}{2}\right)(\exp(\delta_0) - 1).$$

Next, we have

$$\mu_0^p(s) = E\left(W S^{(0)} \mid S^{(0)} = s\right) = s E\left(W^{(0)} \mid S^{(0)} = s\right)$$
$$= s \times \exp\left(\frac{1}{4}\right) s^{\frac{1}{2}} = \exp\left(\frac{1}{4}\right) s^{\frac{3}{2}},$$

and

$$\Delta_P = E\left\{\left(S^{(1)}\right)^{\frac{3}{2}} \exp\left(\frac{1}{4}\right)\right\} - E\left\{\left(S^{(0)}\right)^{\frac{3}{2}} \exp\left(\frac{1}{4}\right)\right\}$$
$$= \exp\left(\frac{5}{2}\right)\left(\frac{3\delta_0}{2} - 1\right).$$

Consequently, in this setting, $\Delta_P > \Delta = \Delta_H$ even though the W has the same distribution in both studies.
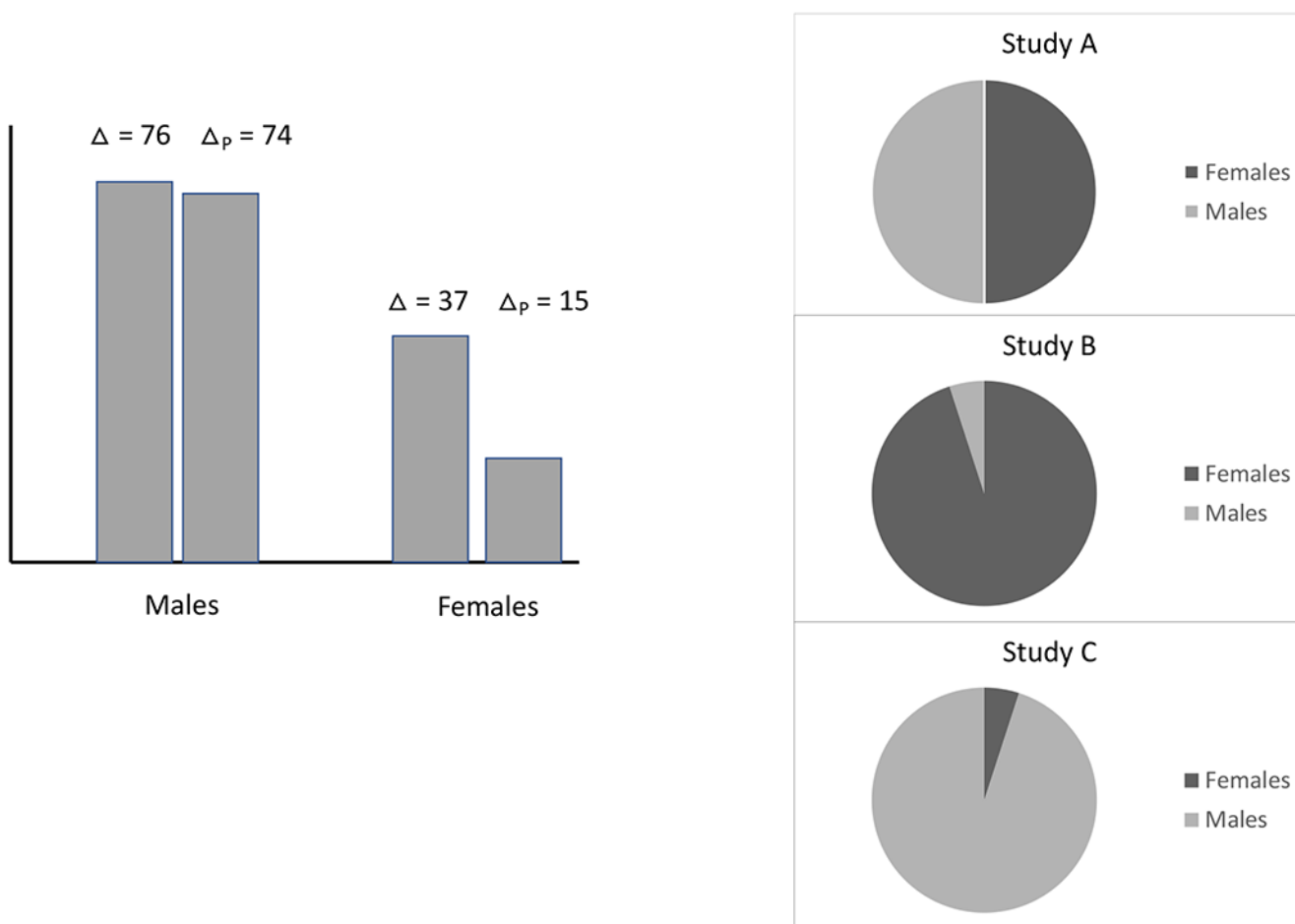

FIGURE 1.
Discrete data example

References

1. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in medicine 1989; 8(4): 431–440. [PubMed: 2727467]

2. Burzykowski T, Molenberghs G, Buyse M. The evaluation of surrogate endpoints. Springer . 2005.

3. Wang Y, Taylor JM. A measure of the proportion of treatment effect explained by a surrogate marker. Biometrics 2002; 58(4): 803–812. [PubMed: 12495134]

4. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. Biometrics 2008; 64(4): 1146–1154. [PubMed: 18363776]

5. Parast L, McDermott MM, Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. Statistics in Medicine 2016; 35(10): 1637–1653. [PubMed: 26631934]

6. Avorn J, Kesselheim AS. Up is down—pharmaceutical industry caution vs. federal acceleration of Covid-19 vaccine approval. New England Journal of Medicine 2020; 383(18): 1706–1708. [PubMed: 32931162]

7. Parast L, Cai T, Tian L. Using a surrogate marker for early testing of a treatment effect. Biometrics 2019; 75(4): 1253–1263. [PubMed: 31009073]

8. Chen X, Hartford A, Zhao J. A model-based approach for simulating adaptive clinical studies with surrogate endpoints used for interim decision-making. Contemporary clinical trials communications 2020; 18: 100562. [PubMed: 32395663]

9. Price BL, Gilbert PB, Laan v dMJ. Estimation of the optimal surrogate based on a randomized trial. Biometrics 2018; 74(4): 1271–1281. [PubMed: 29701875]

10. Athey S, Chetty R, Imbens GW, Kang H. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. tech. rep, National Bureau of Economic Research; 2019.

11. Lin D, Fischl MA, Schoenfeld D. Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. Statistics in medicine 1993; 12(9): 835–842. [PubMed: 8101011]

12. Parast L, Cai T, Tian L. Testing for Heterogeneity in the Utility of a Surrogate Marker. Biometrics, In press 2021.

13. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Statistics in medicine 1992; 11(2): 167–178. [PubMed: 1579756]

14. Parast L, Cai T, Tian L. Evaluating Surrogate Marker Information using Censored Data. Statistics in Medicine 2017; 36(11): 1767–1782. [PubMed: 28088843]

15. VanderWeele TJ. Surrogate measures and consistent surrogates. Biometrics 2013; 69(3): 561–565. [PubMed: 24073861]

16. Lin D, Fleming T, De Gruttola V, others. Estimating the proportion of treatment effect explained by a surrogate marker. Statistics in medicine 1997; 16(13): 1515–1527. [PubMed: 9249922]

17. Eastell R, Barton I, Hannon R, Chines A, Garnero P, Delmas P. Relationship of early changes in bone resorption to the reduction in fracture risk with risedronate. Journal of Bone and Mineral Research 2003; 18(6): 1051–1056. [PubMed: 12817758]

18. Scott D. Multivariate density estimation. Wiley, New York. 1992.

19. Henry K, Erice A, Tierney C, et al. A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS. Journal of acquired immune deficiency syndromes and human retrovirology 1998; 19(4): 339–349. [PubMed: 9833742]

20. Hammer SM, Squires KE, Hughes MD, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. New England Journal of Medicine 1997; 337(11): 725–733. [PubMed: 9287227]

21. ACTG. AIDS Clinical Trial Group: Proposals and Collaboration. https://actgnetwork.org/submit-a-proposal/; 2021.

22. Calmy A, Ford N, Hirschel B, et al. HIV viral load monitoring in resource-limited regions: optional or necessary?. Clinical infectious diseases 2007; 44(1): 128–134. [PubMed: 17143828]

23. Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 2011; 12(2): 270–282. [PubMed: 20876663]

24. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association 2013; 108(502): 527–539. [PubMed: 24058223]

25. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. Statistics in medicine 1994; 13(9): 955–968. [PubMed: 8047747]

26. Pepe MS. Inference using surrogate outcome data and a validation sample. Biometrika 1992; 79(2): 355–365.

27. Elliott MR, Conlon AS, Li Y, Kaciroti N, Taylor JM. Surrogacy marker paradox measures in meta-analytic settings. Biostatistics 2015; 16(2): 400–412. doi: 10.1093/biostatistics/kxu043 [PubMed: 25236906]
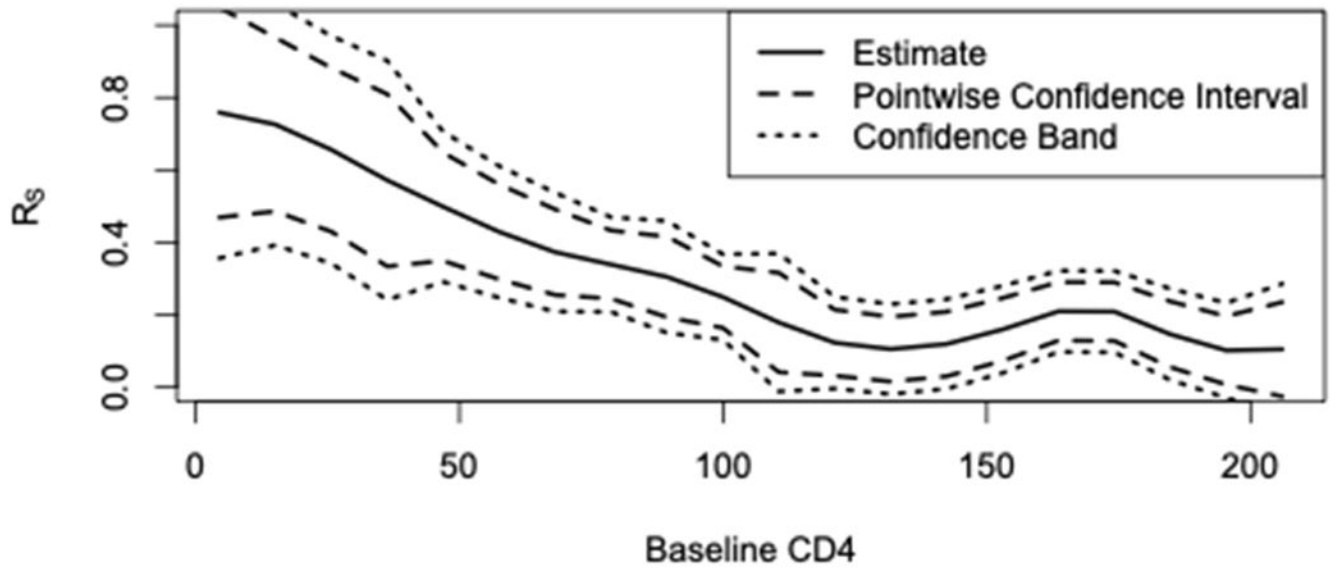
FIGURE 1.

Estimated proportion of the treatment effect on the primary outcome (change in RNA) explained by the treatment effect on the surrogate marker (change in CD4), denoted as R_S, as a function of baseline CD4
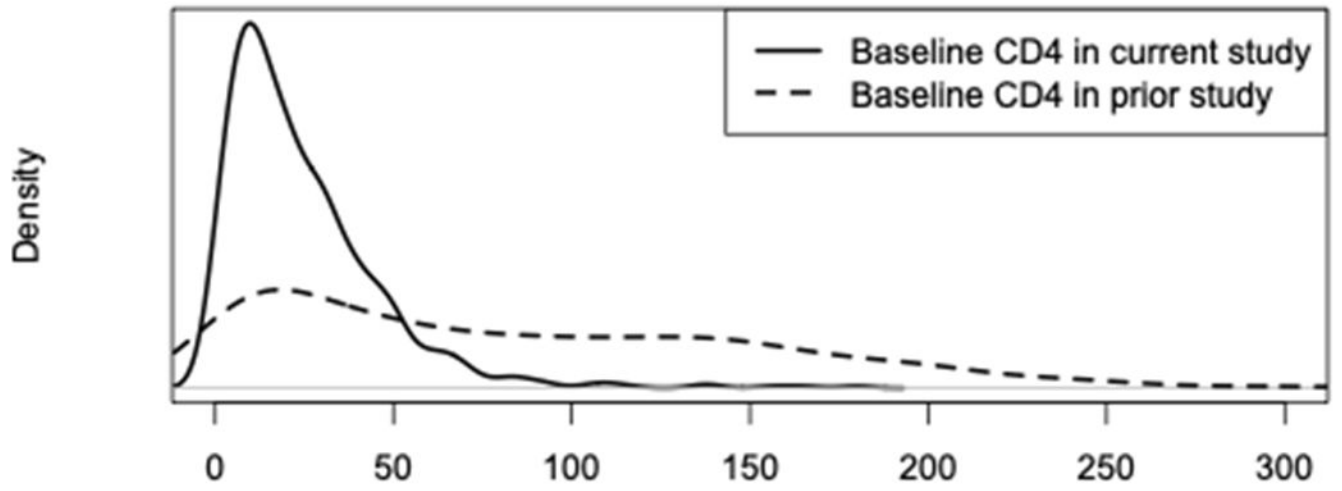
FIGURE 2.
Distribution of baseline CD4 in current study vs. prior study

TABLE 1

Estimation results from the simulation study using the proposed procedure to estimate $\widetilde{\Delta}_H$; note that settings 7 and 8 are null settings with no treatment effect; bias and coverage are examined with respect to $\widetilde{\Delta}_H$ (prior study fixed) and Δ_H; $\widetilde{\mathrm{Bias}}$ = bias with respect to $\widetilde{\Delta}_H$, quantified as $|\widehat{\Delta}_H - \widetilde{\Delta}_H| / \widetilde{\Delta}_H$ except for settings 7 and 8 where it is quantified without dividing by $\widetilde{\Delta}_H$; Bias = bias with respect to Δ_H, quantified as $|\widehat{\Delta}_H - \Delta_H| / \Delta_H$ except for settings 7 and 8 where it is quantified without dividing by the truth; ESE = empirical standard error, ASE = average standard error (average of the square root of the closed form variance estimate), $\widetilde{\mathrm{Cov}}$ = coverage of 95% confidence intervals with respect to $\widetilde{\Delta}_H$; Cov = coverage of 95% confidence intervals with respect to Δ_H

	Estimate	Bias	$\widetilde{\mathrm{Bias}}$	ESE	ASE	Cov	$\widetilde{\mathrm{Cov}}$
Setting 1	6.32	0.07	0.05	1.82	1.79	0.96	0.96
Setting 2	12.53	0.05	0.07	5.39	5.22	0.94	0.94
Setting 3	12.52	0.05	0.07	5.39	5.22	0.94	0.94
Setting 4	14.72	0	0.05	4.12	4.13	0.96	0.95
Setting 5	5.75	0.03	0.04	1.38	1.4	0.95	0.95
Setting 6	12.97	0.01	0.02	1.05	1.27	0.98	0.98
Setting 7	−0.03	0.03	0.16	1.31	1.25	0.94	0.94
Setting 8	−0.03	0.03	0.16	1.31	1.26	0.94	0.94

TABLE 2

Testing results from the simulation study comparing testing results based on the three different testing quantities: $\widehat{\Delta}$ (using the primary outcome, gold standard) vs. $\widehat{\Delta}_P$ (using the surrogate marker, ignoring heterogeneity) vs. $\widehat{\Delta}_H$ (using the surrogate marker, accounting for heterogeneity); ESE = empirical standard error, ASE = average standard error (average of the square root of the closed form variance estimate), Effect size = estimate divided by the estimated standard error (i.e., square root of the closed form variance estimate), Power/Type 1 error = proportion of replications for which the test rejects the null i.e., p-value of the test is < 0.05

	Estimate	ESE	ASE	Effect size	Power
Setting 1					
	14.10	1.64	1.65	8.55	1.00
Δ_P	14.53	3.61	3.65	3.99	0.98
Δ_H	6.32	1.82	1.79	3.62	0.95
Setting 2					
	13.34	5.54	5.42	2.47	0.69
Δ_P	7.64	3.38	3.31	2.31	0.64
Δ_H	12.53	5.39	5.22	2.39	0.67
Setting 3					
	13.34	5.54	5.42	2.47	0.69
Δ_P	6.00	2.81	2.76	2.18	0.58
Δ_H	12.52	5.39	5.22	2.39	0.67
Setting 4					
	19.12	5.17	5.20	3.68	0.96
Δ_P	14.64	3.66	3.66	4.01	0.98
Δ_H	14.72	4.12	4.13	3.56	0.95
Setting 5					
	13.90	1.64	1.65	8.43	1.00
Δ_P	5.77	1.38	1.38	4.18	0.99
Δ_H	5.75	1.38	1.40	4.09	0.99

Setting 1					
	Estimate	**ESE**	**ASE**	**Effect size**	**Power**
Setting 6					
	Estimate	**ESE**	**ASE**	**Effect size**	**Power**
	33.70	1.61	1.60	21.08	1.00
Δ_P	39.12	3.51	3.50	11.18	1.00
Δ_H	12.97	1.05	1.27	10.23	1.00
Setting 7					
	Estimate	**ESE**	**ASE**	**Effect size**	**Type 1 error**
	−0.05	1.39	1.35	−0.04	0.06
Δ_P	−0.03	1.31	1.27	−0.02	0.06
Δ_H	−0.03	1.31	1.25	−0.02	0.06
Setting 8					
	Estimate	**ESE**	**ASE**	**Effect size**	**Type 1 error**
	−0.05	1.37	1.33	−0.04	0.06
Δ_P	−0.03	1.31	1.27	−0.02	0.06
Δ_H	−0.03	1.31	1.26	−0.02	0.06