Statistics in Medicine

WILEY

Statistics in Medicine

TUTORIAL IN BIOSTATISTICS

Surrogate Marker Evaluation: A Tutorial Using R

Layla Parast

Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, USA

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

Received: 22 November 2024 | **Revised:** 11 February 2025 | **Accepted:** 23 February 2025

Funding: This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases, R01DK118354.

Keywords: biomarker | censoring | implementation | non-parametric | surrogate | treatment effect

ABSTRACT

The practice of using a surrogate marker to replace a primary outcome in clinical studies has become widespread. Typically, the primary outcome requires long-term patient follow-up, is expensive, or is invasive or burdensome for patients to measure, while the surrogate marker is not (or less so). Of course, a surrogate marker must be validated before it should be used to make a decision about the effectiveness of a treatment. There has been a tremendous amount of statistical and clinical research focused on evaluating and validating surrogate markers over the past 35 years. Although there is ongoing debate over the optimal evaluation method, the development of new approaches and insights has greatly enriched the field. In this tutorial, we describe available statistical frameworks for evaluating a surrogate marker and specifically focus on the practical implementation of the proportion of treatment effect explained framework. We consider both uncensored and censored outcomes, parametric and non-parametric estimation, evaluating multiple surrogates, heterogeneity in the utility of the surrogate marker, surrogate evaluation from a prediction perspective, and the surrogate paradox. We include R code to implement these procedures with a follow-along R markdown. We close with a discussion on open problems in this research area, particularly in terms of using the surrogate marker to test for treatment in a future study, which is the ultimate goal of surrogate marker evaluation.

1 | Introduction

Randomized clinical trials are the gold standard for evaluating the effect of new treatments or interventions on patient outcomes. Such trials are often focused on a primary outcome that may require long-term patient follow-up or is expensive or burdensome to measure. In this case, there is tremendous interest in using a surrogate marker or surrogate outcome to evaluate the treatment effect, instead of the primary outcome, in order to make decisions about the treatment more quickly or with less cost.

In the United States, it is indeed possible to receive drug approval based on demonstrated effectiveness on a surrogate marker via the Food and Drug Administration's (FDA) Accelerated Approval Program [1]. For example, the FDA originally approved a new drug, Jardiance, for people with Type 2 diabetes based on results showing that the drug lowered blood glucose levels, a surrogate marker [2]. Two years later, results showed that the drug indeed reduced the risk of cardiovascular death in these patients [3]. A less positive example is the Alzheimer's drug aducanumab which was originally approved based on results showing that it reduced visible plaque in the brain, but was later found to have no effects on patient outcomes [4–7]. The impact of this controversial drug was dramatic. Not surprisingly, stock prices for Biogen, the company that made the drug, were highly sensitive to the approval trajectory, surging when the

initial results came up, declining when the outcome results were made public, increasing again when the FDA announced they would not pull the drug, and declining again when Medicare announced they would restrict coverage for the drug and private insurers followed suit [8–11]. Biogen subsequently voluntarily removed the drug from the market [12].

While it is clearly problematic if a drug has a positive effect on a surrogate marker but no effect on the primary outcome, it is devastating if a drug has a positive effect on a surrogate marker but a negative effect on the primary outcome, particularly if the primary outcome is death. The most widely cited example of this is anti-arrhythmia drugs which were approved based on results showing that they lowered the incidence of arrhythmias but were then found to dramatically increase cardiac deaths [13–15]. It has been estimated that the approval of these drugs resulted in tens of thousands of deaths in the United States [14]. Importantly, this approval process was not at all based on any formal statistical evaluation of arrhythmias as a surrogate for cardiovascular events or death. Instead, the acceptance of arrhythmias as a surrogate was based on clinical expertise and prior studies examining a single outcome. That is, prior studies showed that the drug reduced arrhythmias and separate studies had shown that reduced arrhythmias result in fewer cardiovascular deaths. No randomized study had tested whether these drugs measured both arrhythmias and cardiovascular deaths until finally, the Cardiac Arrhythmia Suppression Trial (CAST) was conducted [16–18]. The CAST study was stopped early due to overwhelming evidence that these drugs increased cardiovascular death [17]. Clearly, formal statistical validation of a potential surrogate marker is extremely important before the surrogate marker is used to make decisions about a treatment effect.

Though the FDA does provide a definition of a surrogate maker, the statistical implications of the definition are lacking. Specifically, they define a surrogate as an outcome or marker that measures "a therapeutic effect that is considered reasonably likely to predict the clinical benefit of a drug, such as an effect on irreversible morbidity and mortality." [1] Fortunately, Ross Prentice [19] formalized a statistical definition by proposing a criterion for a valid surrogate marker requiring that a test for treatment effect on the surrogate marker must also be a valid test for treatment effect on the primary outcome of interest. While there is some controversy around this criterion and there is still no single agreed-upon method to validate a surrogate, Prentice's work spurred 35 years of statistical methodological work aimed at validating surrogate markers [20].

In this tutorial, we describe available statistical frameworks for evaluating surrogate markers, with an emphasis on the practical implementation of the proportion of treatment effect explained framework. It is important that we highlight our focus on "statistical" frameworks; it is implicit that a surrogate marker should only undergo statistical evaluation if there is sufficient biological evidence supporting its potential relevance. We consider both uncensored and censored outcomes, parametric and non-parametric estimation, evaluating multiple surrogates, heterogeneity in the utility of the surrogate marker, and surrogate evaluation from a prediction perspective. Throughout, we let Y denote the primary outcome, S denote the surrogate marker, and Z denote the treatment indicator where $Z \in \{0, 1\}$ (i.e., treatment vs. control) and treatment is assumed to be randomized unless otherwise noted. Without loss of generality, we assume that higher values of Y and S are "better". The observed data consists of $\{Y_i, S_i, Z_i\}$ for each individual i. This article is structured as follows. In Section 2, we introduce notation and describe and compare available frameworks. In Section 3, we specifically describe and implement the proportion of treatment effect framework, and in Section 4, we expand to multiple surrogates, surrogates measured with error, censored outcomes, heterogeneity in the utility of a surrogate, surrogate evaluation from a prediction perspective, and the surrogate paradox. We include R code to implement these procedures with a follow-along R markdown available at: https://github.com/laylaparast/SIMtutorial; all referenced packages are available on CRAN [21]. Throughout, both a heuristic explanation and the mathematical details of the methods are provided, with the understanding that a full grasp of the mathematical intricacies is not strictly required. Finally, in Section 5, we discuss open problems in this research area, particularly in terms of using the surrogate marker to test for treatment in a future study and describe the connection between surrogate assessment and mediation analysis. It is important to emphasize here that the primary goal in surrogate evaluation is to ultimately replace the primary outcome with the surrogate to make inferences on the treatment effect in a future trial. Additional recent reviews on this topic that may be helpful in conjunction with this tutorial include Elliott (2023) [22], discussed below, Parast et al. (2024) [23], which is geared towards a clinical audience, and Gilbert et al. (2024) [24], which focuses on the vaccine setting.

2 | Available Frameworks for Surrogate Validation

A recent review by Elliott (2023) [22] describes available frameworks for surrogate validation in detail. For the purpose of this tutorial, we describe three frameworks below: The proportion of treatment effect explained framework, the principal stratification framework, and the meta-analytic framework. Certainly, other frameworks and approaches exist but are not covered here including methods motivated by information-theoretic concepts [25, 26]. Previous work has discussed the links between these frameworks including, but not limited to, Alonso et al. (2004), [27] Conlon et al. (2017), [28] and Stijven et al. (2024) [29].

2.1 | Proportion of Treatment Effect Explained

First, we describe the proportion of treatment effect explained (PTE), which we combine with the direct and indirect framework, for reasons that will shortly become clear. The full terminology is the proportion of the treatment effect on the primary outcome that is explained by the treatment effect on the surrogate, but this is often shortened to simply PTE. Motivated by the work of Prentice,

Freedman et al. (1992) [30] proposed to evaluate a surrogate marker by defining and estimating the PTE via specifying two regression models, for example:

$$E(Y|Z) = \beta_0 + \beta_1 Z \tag{1}$$

$$E(Y|Z,S) = \beta_0^* + \beta_1^* Z + \beta_2 S \tag{2}$$

where the PTE is defined as $R_F = 1 - \beta_1^*/\beta_1$ and estimated by plugging in the corresponding regression estimates. Intuitively, the idea is that if one fits model (1) and observes a large treatment effect (large β_1) and then fits model (2), essentially adding in the surrogate marker, and the treatment effect is now small or close to 0 (small β_1^*), then R_F will be close to 1, indicating that the surrogate is capturing the effect of the treatment on Y. This approach is extremely appealing and easy to implement. However, numerous studies have pointed out problems with this approach, one of which is that it relies on both models being correctly specified. Notably, not only is the estimate dependent on correct specification, but the definition of the quantity R_F itself relies on correct specification. In a survival setting, where the outcome is a censored time-to-event outcome, if Cox proportional hazards models are used for models (1) and (2), it is actually impossible for both models to hold simultaneously; see Lin et al. (1997) [31] for more details.

As an alternative, Wang and Taylor (2002) [32] proposed a different approach that aims to capture the same idea with a model-free definition. To introduce this, we will use potential outcomes notation where each person has a potential $\{Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\}$ where $Y^{(g)}$ is the outcome when $Z = g$ and $S^{(g)}$ is the surrogate when $Z = g$. They propose to quantify the PTE using contrasts between the actual treatment effect on Y, defined as:

$$\Delta = E(Y^{(1)} - Y^{(0)})$$

and the *residual treatment effect* on Y defined as

$$\begin{aligned}\Delta_S &= E_{S^{(0)}}[E(Y^{(1)} - Y^{(0)}|S^{(1)} = S^{(0)} = s)] \\ &= \int E(Y^{(1)} - Y^{(0)}|S^{(1)} = S^{(0)} = s) dF_{S^{(0)}}(s)\end{aligned}$$

where $F_{S^{(0)}}$ is the cumulative distribution function of $S^{(0)}$. The residual treatment effect can be interpreted as the "leftover" treatment effect on Y, after accounting for the treatment effect on S. That is, it is the hypothetical treatment effect on Y if the distribution of the surrogate in both groups looked like the distribution of the surrogate in the control group. Importantly, the definition of Δ_S uses the distribution of $S^{(0)}$, but in theory, one can select the distribution of $S^{(1)}$ or some combination of the two. The PTE is then defined as:

$$R_W = \frac{\Delta - \Delta_S}{\Delta} = 1 - \frac{\Delta_S}{\Delta}$$

where $\Delta - \Delta_S$ is the treatment effect explained by S. This decomposition of Δ into Δ_S and $\Delta - \Delta_S$ parallels the direct/indirect framework of Robins and Greenland (1992) [33]. Ideally, this quantity is between 0 and 1, with values close to 0 indicating a poor surrogate (not capturing the treatment effect) and values close to 1 indicating a good surrogate (able to capture the treatment effect). However, R_W is only guaranteed to be within $[0, 1]$ under certain assumptions, detailed in Section 4.6. This construction of R_W highlights a key challenge in surrogate evaluation: When the overall treatment effect, Δ, is small, identifying a surrogate becomes inherently difficult. In this particular framework, a small treatment effect implies that finding a surrogate capable of explaining a large proportion of such a small effect will be highly challenging. Much has been and can be argued about this definition of Δ_S and R_W in terms of reference distribution, types of indirect effect, identifiability, causal mechanisms, and more; however, we omit these arguments here and instead refer interested readers to existing work [24, 28, 29, 34–37].

Unlike R_F, the definition R_W does not involve any model specification and thus, is model-free. In terms of estimation of R_W, there are both parametric and non-parametric options available, which we detail in Section 3. Overall, this framework offers a single number, the PTE, that aims to quantify the strength of the surrogate marker with respect to capturing the treatment effect on Y. While there is no agreed-upon threshold for a "good enough" surrogate, some have proposed using 0.5 or 0.75 as a threshold for either the point estimate or, more strictly, for the lower bound of the confidence interval [30, 31, 38]. Software to implement this approach is available in the R package `Rsurrogate` on CRAN, which is described in more detail in Section 3.

2.2 | Principal Stratification

The second framework is one based on principal stratification. Principal stratification is a causal inference method proposed by Frangakis and Rubin (2002) [39] and generally developed for a setting where one aims to condition jointly on the potential outcomes of two variables. Surrogate marker evaluation is one application of principal stratification, developed further by Gilbert and Hudgens (2008) [40]. The framework is based on principal effects which are defined as comparisons of potential outcomes within a principal stratum:

$$E\left(Y^{(1)} - Y^{(0)}|(S^{(1)}, S^{(0)}) = (s_1, s_0)\right)$$

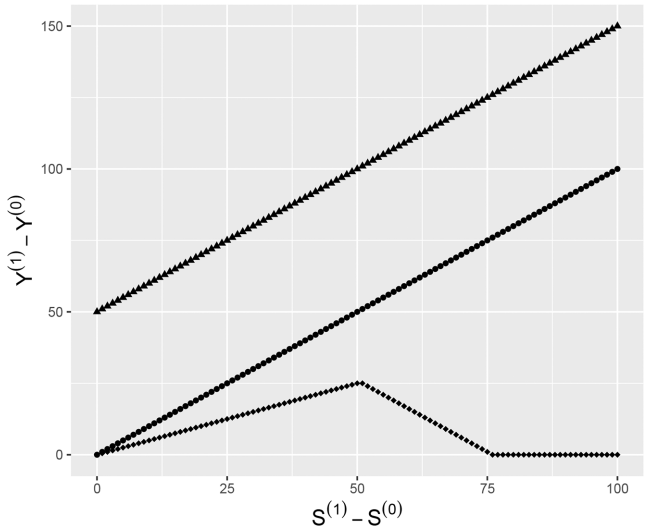

FIGURE 1 | Simplified illustration of the Causal Effect Predictiveness (CEP) Curve; a CEP Curve where (a) both average causal necessity and average causal sufficiency hold (black circles), (b) average causal necessity holds but average causal sufficiency does not hold (black diamonds), and (c) neither average causal necessity and average causal sufficiency hold (black triangles).

In this framework, for S to be a "principal surrogate", we must have (1) Average Causal Necessity (ACN):

$$E(Y^{(1)}|S^{(1)} = S^{(0)} = s) = E(Y^{(0)}|S^{(1)} = S^{(0)} = s) \quad \text{for all} \quad s$$

and (2) Average Causal Sufficiency (ACS):

$$E(Y^{(1)}|S^{(1)} = s_1, S^{(0)} = s_0) \neq E(Y^{(0)}|S^{(1)} = s_1, S^{(0)} = s_0) \text{ for all } \quad |s_1 - s_0| > C$$

Essentially, ACN means that if $S^{(1)}$ and $S^{(0)}$ are the same, then $Y^{(1)}$ and $Y^{(0)}$ should also be the same; and ACS means that if $S^{(1)}$ and $S^{(0)}$ are different, then $Y^{(1)}$ and $Y^{(0)}$ should also be different. Of course, the structure of the ACN and ACS conditions look quite similar to the components of the Δ_S quantity within the PTE framework, and indeed, prior work has demonstrated the links between these two frameworks [28, 29]. For example, Stijven et al. (2024) [29] show that if ACN holds, then $\Delta_S = 0$. The concepts of ACN and ACS can be visualized via the Causal Effect Predictiveness (CEP) Curve which is a curve describing the relationship between $S^{(1)} - S^{(0)}$ and $Y^{(1)} - Y^{(0)}$. We describe this using three simple hypothetical examples, shown in Figure 1. One example reflects a case where both average causal necessity and average causal sufficiency hold (black circles), which can be seen by the CEP curve going through the origin and having a positive slope. Another example (black diamonds) reflects a case where average causal necessity holds (goes through the origin) but average causal sufficiency does not hold (there are values where $|S_1 - S_0| > C$ but $Y^{(1)} = Y^{(0)}$), and a last example (black triangles) where average causal necessity does not hold (does not go through the origin) but average causal sufficiency holds.

In reality, these quantities are not identifiable because we generally do not have both $S^{(1)}$ and $S^{(0)}$ (and $Y^{(1)}$ and $Y^{(0)}$) for the same individual. However, there are some exceptions such as within the crossover trial design discussed in Gabriel and Follman (2016) [41] and in certain vaccine settings where S measures immune response and there is no possibility of an immune response if given a placebo, that is, $S^{(0)} = 0$ for all individuals, referred to as the constant-biomarker setting, discussed in Gilbert and Hudgens (2008) [40]. To estimate the CEP Curve, identifiability assumptions are typically needed, and in one way or another, a parametric assumption is imposed. For example, Huang and Gilbert (2011) [42] propose an approach that uses baseline covariates, W, to predict unobserved S and then a generalized linear model describing the dependence of Y on S and Z with a specified parametric link function. Other estimation approaches within the principal stratification framework include Bayesian procedures, and useful extensions to other settings including survival and longitudinal surrogate settings [24, 41, 43–48].

2.3 | Meta-Analytic

The third framework is the meta-analytic framework and is applicable when multiple studies are available [49–51]. In this setting, the observed data consists of $\{Y_{ij}, S_{ij}, Z_{ij}\}$ for individual i in trial/study j where $j = 1, \ldots, J$ with J being the number of trials. This framework, which considers random trial-level intercepts and subject-level correlations, generally relies on the following bivariate model specification:

$$\begin{cases} Y_{ij} = \nu_{Yi} + \theta_i Z_{ij} + \epsilon_{Yij} \\ S_{ij} = \nu_{Si} + \gamma_i Z_{ij} + \epsilon_{Sij} \end{cases} \tag{3}$$

where ν_{Yi} and ν_{Si} are trial-specific intercepts, θ_i and γ_i are trial-specific treatment effects, $\{\nu_{Yi}, \nu_{Si}, \theta_i, \gamma_i\}$ is assumed to follow a normal distribution, and ϵ_{Yij} and ϵ_{Sij} are correlated error terms assumed to have a bivariate normal distribution such that

$$\begin{pmatrix} \epsilon_{Yij} \\ \epsilon_{Sij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{YY} & \sigma_{SY} \\ \sigma_{SY} & \sigma_{SS} \end{pmatrix} \right)$$

The assumed normality means that one can essentially derive the expected treated effect on Y conditional on the treatment effect on S, which is of course our ultimate goal when using a surrogate marker in a future trial. The assumed normality results in various terms working out quite nicely with some algebra. (Even if Y and S are not normally distributed, it may be possible to determine an appropriate transformation to make this assumption reasonable.) Surrogacy is typically quantified using the proportion of variance in the total effect explained by the trial-level random effects associated with the surrogate, denoted as R^2_{trial}, which is a function of the various variance matrix quantities. Larger values of R^2_{trial} reflect a stronger surrogate. An advantage of this approach is the lack of needed assumptions about relationships between unobserved potential outcomes which are needed in the PTE and principal stratification frameworks. Disadvantages of this approach include its dependence on the assumption of a normal distribution, which is often unrealistic in practice, and, more critically, the clear requirement for data from multiple trials. While some may consider the meta-analytic approach the gold standard, its main limitation is that the necessary data are rarely available. In many cases, only a single study is available to validate a surrogate marker. Software to implement this approach is available in the R package `Surrogate` on CRAN. Alternative estimation approaches, useful extensions to other settings including survival settings, as well as methods that utilize principal stratification within a meta-analytic setting exist [26, 52–56].

3 | Proportion of Treatment Effect Explained: Details and Implementation

For the remainder of this tutorial, we focus on the proportion of the treatment effect framework and specifically discuss estimation, implementation in R, and interpretation. We emphasize this framework primarily because of the availability of R packages for its implementation, as well as its clinical appeal and usage by applied researchers. This section covers material proposed in Freedman et al. (1992), [30], Wang and Taylor (2002), [32], and Parast et al. (2016) [57]; methodological details and justification can be found therein. We use functions available in the `Rsurrogate` package on CRAN. As mentioned earlier, we include R code to implement these procedures within this paper, and also provide a follow-along R markdown document.

For illustration, we use a hypothetical dataset `d_example` available in the `Rsurrogate` package which is a list with 500 observations from a control group and 500 observations from a treatment group; list elements `y1` and `y0` are the primary outcomes for the treatment and control observations, respectively, and `s1.a` and `s0.a` are the surrogate marker measurements for the treatment and control observations, respectively. First, using the function `R.s.estimate`, we estimate R_F described in Section 2.1, which relies on the correct specification of models (1) and (2) and is calculated by simply using the estimates from these fitted regression models. The function `R.s.estimate` is the main function of this package. The primary arguments are `sone` and `szero`, the vectors of surrogate marker measurements in the treated group and the control group, respectively, and `yone` and `yzero`, the vectors of primary outcomes in the treated group and the control group, respectively. The argument `type` is the type of estimation the user would like with the options being "freedman", which is the regression approach described above, and "model" and "robust" which we describe below. The default is "robust". The user can ask for the variance and confidence intervals by including `var=TRUE` and `conf.int=TRUE`, respectively. As described in Parast et al. (2016) [57], there are multiple ways one can construct a confidence interval. This function provides three intervals: One based on a normal approximation (`conf.int.normal.R.s`), one based on sample percentiles from a resampled distribution (`conf.int.quantile.R.s`), and one based on constructing an interval for a ratio via Fieller's theorem (`conf.int.fieller.R.s`). In practice, we have found these to perform similarly and recommend using `conf.int.quantile.R.s`.

In the code below, we install and load the package, take a look at the data, estimate R_F, and obtain variance estimates and confidence intervals. Note that we set a seed before the variance estimation command; this is to ensure reproducibility since these estimates are resampling-based.

```
#install the package from CRAN
install.packages("Rsurrogate")

#load the R package
library(Rsurrogate)

#take a look at the data
data(d_example)
names(d_example)
```

```
## [1] "s1.a" "s1.b" "s1.c" "y1"   "s0.a" "s0.b" "s0.c" "y0"

#Estimate the PTE using the Freedman approach
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone = d_example$y1,
yzero = d_example$y0, type = "freedman")

## $R.s
## [1] 0.2161019

#ask for variance and confidence intervals
#set.seed(1)
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone = d_example$y1,
yzero = d_example$y0, type = "freedman", var = TRUE, conf.int = TRUE)

## $R.s
## [1] 0.2161019
##
## $R.s.var
## [1] 0.0005117975
##
## $conf.int.normal.R.s
## [1] 0.1717609 0.2604429
##
## $conf.int.quantile.R.s
## [1] 0.1718017 0.2591348
##
## $conf.int.fieller.R.s
## [1] 0.1705914 0.2597199
```

This code results in an estimate of 0.22 with a 95% confidence interval (0.17, 0.26). That is, we estimate that the proportion of the treatment effect explained by the surrogate marker is 0.22 (95% CI: 0.17,0.26). As this is far below a threshold of 0.5 (and 0.75), we would conclude that this is not a reasonable surrogate. Of course, whether this interpretation is valid relies upon the validity of the parametric assumptions. Namely, if the models (1) and (2) are incorrectly specified, then this quantity is not meaningful.

We move on now to the model-free definition of Wang and Taylor (2002) [32] described in Section 2.1, R_W. Wang and Taylor (2002) [32] propose an estimation approach for R_W which is still parametric, but can be more flexible than the models imposed by R_F. For example, one could specify the following:

$$E(Y^{(0)}|S^{(0)} = s) = \gamma_0 + \gamma_1 s \tag{4}$$

$$E(Y^{(1)}|S^{(1)} = s) = \gamma_2 + \gamma_3 s \tag{5}$$

It can be shown algebraically that under these models, $\Delta = (\gamma_2 - \gamma_0) + \gamma_1\alpha_0 + \gamma_3\alpha_1$ and $\Delta_S = (\gamma_2 - \gamma_0) + (\gamma_3 - \gamma_1)\alpha_0$ where $\alpha_0 = E(S^{(0)})$, and $\alpha_1 = E(S^{(1)})$. (Note that if $\gamma_1 = \gamma_3$ then it can be shown that $R_W = R_F$.) Thus, R_W can be estimated by plugging in the regression estimates from fitting models (4) and (5) and the average of S in each group as estimates of α_0 and α_1. This can be implemented as follows:

```
#Estimate the PTE using Wang and Taylor approach
#notice that now we are given delta, delta.s, and R.s
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone=d_example$y1,
yzero=d_example$y0, type = "model")

## $delta
## [1] 7.727727
##
## $delta.s
## [1] 6.068396
##
```

```
## $R.s
## [1] 0.2147244

#ask for variance and confidence intervals
set.seed(1)
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone = d_example$y1,
yzero = d_example$y0, type = "model", var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.2147244
##
## $conf.int.quantile.R.s
## [1] 0.1699794 0.2571792
```

Note that the only change in the function call is to change `type` to "model". This type produces more output as now the function will return not only an estimate of R_W, but also of Δ and Δ_S as well as the associated variance estimates and confidence intervals (when those arguments are set to `TRUE`). For Δ and Δ_S, the function will provide only the first two types of confidence intervals as the third type based on Fieller's theorem is only relevant for a ratio (only R_W). Our results are similar in that the estimated proportion of the treatment effect explained is 0.21 (95% CI: 0.17, 0.26).

Lastly, we move on to the non-parametric estimation approach which relies on kernel smoothing and is described in detail in Parast et al. (2016) [57]. Let the observed data be denoted as $\{Y_{1i}, S_{1i}\}$ for $i = 1, \ldots, n_1$ individuals in the treatment group and $\{Y_{0j}, S_{0j}\}$ for $j = 1, \ldots, n_0$ individuals in the control group. First, Δ is estimated simply as the standard average treatment effect appropriate in a randomized study:

$$\widehat{\Delta} = \sum_{i=1}^{n_1} Y_{1i} - \sum_{j=1}^{n_0} Y_{0j}$$

The residual treatment effect, Δ_S is estimated using the Nadaraya-Watson conditional mean estimator [58, 59]:

$$\begin{aligned} \widehat{\Delta}_S &= \sum_{j=1}^{n_0} \widehat{\mu}_1(S_{0j}) - \sum_{j=1}^{n_0} Y_{0j} \\ \text{where } \widehat{\mu}_1(s) &= \frac{\sum_{i=1}^{n_1} K_h(S_{1i} - s) Y_{1i}}{\sum_{i=1}^{n_1} K_h(S_{1i} - s)} \end{aligned} \tag{6}$$

is the non-parametric estimate of $\mu_1(s) = E(Y^{(1)}|S^{(1)} = s)$, $K(\cdot)$ is a smooth symmetric density function, $K_h(\cdot) = K(\cdot/h)/h$, and h is a specified bandwidth. Let's take a look at where this estimator came from: Taking Δ_S defined earlier, an identifiability assumption allows us to break up the conditional expectation inside the integral into two pieces:

$$\begin{aligned} \Delta_S &= \int E(Y^{(1)} - Y^{(0)}|S^{(1)} = S^{(0)} = s) dF_{S^{(0)}}(s) \\ &= \int E(Y^{(1)}|S^{(1)} = s) dF_{S^{(0)}}(s) - \int E(Y^{(0)}|S^{(0)} = s) dF_{S^{(0)}}(s) \\ &= \int E(Y^{(1)}|S^{(1)} = s) dF_{S^{(0)}}(s) - E(Y^{(0)}) \\ &= \int \mu_1(s) dF_{S^{(0)}}(s) - E(Y^{(0)}) \end{aligned} \tag{7}$$

The first term in $\widehat{\Delta}_S$ is simply the empirical estimate of the integral term in (7), where we construct the conditional mean estimate, $\widehat{\mu}_1(s)$, using the treatment group, and apply the estimate to the control group. The second term, of course, is simply the empirical estimate of $E(Y^{(0)})$. The needed identifiability assumption is:

$$Y^{(1)} \perp S^{(0)}|S^{(1)} \text{ and } Y^{(0)} \perp S^{(1)}|S^{(0)}$$

Notably, while this approach is completely non-parametric and does not involve any parametric assumptions, it does still impose this (untestable) identifiability assumption. In the end, the estimate of the PTE is defined as $\widehat{R}_S = 1 - \widehat{\Delta}_S/\widehat{\Delta}$. This can be implemented as follows:

```
#Estimate the PTE using the non-parametric approach
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone = d_example$y1,
yzero = d_example$y0, type = "robust")

## $delta
## [1] 7.727727
##
## $delta.s
## [1] 6.322881
##
## $R.s
## [1] 0.181793

#ask for variance and confidence intervals
set.seed(1)
R.s.estimate(sone = d_example$s1.a, szero = d_example$s0.a, yone = d_example$y1,
yzero = d_example$y0, type = "robust", var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.181793
##
## $conf.int.quantile.R.s
## [1] 0.1226421 0.2378033
```

Again, note that the only change in the function call is to change `type` to "robust". It should come as no surprise that this approach is more computationally intensive given the kernel smoothing aspect. Our results are similar, though a bit lower, in that the estimated PTE is 0.18 (95% CI: 0.12, 0.24). Two implementation details are worth discussing here. First, the bandwidth h is automatically selected by the function by setting $h = h^* n_1^{-0.25}$ where h^* is obtained using the `bw.nrd` function applied to $\{S_{1i}\}, i = 1, \ldots, n_1$. This is because the asymptotic derivations require that $h = O(n_1^{-\delta})$ with $\delta \in (1/4, 1/2)$; the `bw.nrd` function results in a bandwidth, h^*, of order $n_1^{-1/5}$, thus, we must take $h = h^* n_1^{-c_0}$ where $c_0 \in (1/20, 3/10)$ and the 0.25 value that is used within the function is within this interval. Second, if the support of the observed data $S_{0j}, j = 1, \ldots, n_0$ is outside of the support of $S_{1i}, i = 1, \ldots, n_1$, extrapolation of some kind for $\hat{\mu}_1(s)$ will be needed. This is implemented in the function by taking the value at the closest data extreme, similar to the `approx` function in the `stats R` package. The user can specify if they would like extrapolation to be used via the `extrapolate` argument which can be set to either `TRUE` or `FALSE`; the default for the function is `TRUE`. When the function does implement extrapolation, a message is given to the user. We can recreate this as follows:

```
#take a look at the current ranges
range(d_example$s1.a)
## [1] -1.084469 12.387819

range(d_example$s0.a)
## [1] 1.634990 7.906001

#create an example where there will be a support problem
s1.a.temp = d_example$s1.a[d_example$s1.a < 6]
y1.temp = d_example$y1[d_example$s1.a < 6]

#see resulting warning message
R.s.estimate(sone = s1.a.temp, szero = d_example$s0.a, yone = y1.temp,
yzero = d_example$y0, type = "robust")

## [1] "Warning: Observed supports do not appear equal, may need to consider a transforma-
tion or extrapolation"

## $delta
## [1] 4.981686
##
## $delta.s
```

```
## [1] 6.188534
##
## $R.s
## [1] -0.242257
```

The `R.s.estimate` function also includes an argument `number` that corresponds to the number of surrogate markers being evaluated which can be either "single" or "multiple". In this section, we have focused on evaluating a single surrogate marker and have used the default of `number = "single"`. In the following section, we illustrate the "multiple" option as well as other functions within the `Rsurrogate` package.

4 | Advanced Settings and Considerations in Surrogate Evaluation

4.1 | Multiple Surrogates

In some settings, one may be interested in estimating the proportion of the treatment effect explained by not only a single potential surrogate marker, but rather a set of potential surrogate markers. For example, in a diabetes prevention clinical trial setting, we may wish to evaluate the surrogacy of the 6-month change in fasting plasma glucose, cholesterol, and blood pressure together. The methods described in Section 3 can be extended to handle multiple surrogates and are implemented within the `R.s.estimate` function. The methodological details and justification can be found in Parast et al. (2016) [57]. Here, we briefly describe and illustrate these extended methods. For the Freedman approach, one can simply include each of the surrogates as explanatory variables in the model for $E(Y|S)$, that is, model (2). It would still be the case that the PTE is defined as $R_F = 1 - \beta_1^*/\beta_1$. In the `d_example` dataset, there are three surrogate markers, `s1.a, s1.b, s1.c`, for each individual. Below, we use the `R.s.estimate` function to estimate R_F for this set of surrogates.

```
# Estimate the PTE of multiple markers using the Freedman approach
set.seed(1)
R.s.estimate(yone = d_example$y1, yzero = d_example$y0, sone =
cbind(d_example$s1.a, d_example$s1.b, d_example$s1.c), szero =
cbind(d_example$s0.a, d_example$s0.b, d_example$s0.c), number =
"multiple", type = "freedman", var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.4725655
##
## $conf.int.quantile.R.s
## [1] 0.4217517 0.5150648
```

Importantly, the arguments `sone` and `szero` should be matrices where each column represents a surrogate marker, and we have used the option `number="multiple"` to denote that we want the PTE for a set of multiple markers. Results show that the estimated PTE is 0.47 (95% CI: 0.42, 0.52) meaning that this set of surrogate markers captures 47% of the treatment effect on the outcome, if models (1) and (2) are correctly specified. Similarly, the estimation approach of Wang and Taylor can be modified by extending models (4) and (5) to include each of the surrogates as explanatory variables and following through with some algebra. This can be implemented using:

```
# Estimate the PTE of multiple markers using the Wang & Taylor approach
set.seed(1)
R.s.estimate(yone=d_example$y1, yzero=d_example$y0, sone=
cbind(d_example$s1.a,d_example$s1.b, d_example$s1.c), szero=
cbind(d_example$s0.a, d_example$s0.b, d_example$s0.c), number =
"multiple", type = "model", var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.4829253
##
## $conf.int.quantile.R.s
## [1] 0.4338298 0.5261389
```

Results show that the estimated PTE is 0.48 (95% CI: 0.43, 0.53), similar to the Freedman approach. Lastly, the non-parametric approach can be extended with one additional step wherein a working model is used to reduce the dimension of the set of surrogates to a single score. Specifically, a working model such as

$$E(Y^{(1)}|S^{(1)} = s) = g(\theta' s) \tag{8}$$

where $g(\cdot)$ is a pre-specified monotone increasing function, is used for dimension reduction, and $\widehat{Q}$ is defined as $\widehat{Q} = \widehat{\theta}' S$, a single score. This is referred to as a "working model" because even if the model (8) is misspecified, the estimator $\widehat{\theta}$ still converges to some deterministic limit θ_0. Thus, $Q = \theta_0' S$ is considered the pseudo-surrogate marker of interest and the focus is shifted to estimate the PTE of Q, a particular summary of S. With this extra step, we are now back to a single surrogate setting and can apply the kernel-based estimator of (6) to estimate the PTE of $\widehat{Q}$. This two-step approach can be implemented using:

```
# Estimate the PTE of multiple markers using the robust approach
set.seed(1)
R.s.estimate(yone=d_example$y1, yzero=d_example$y0, sone=
cbind(d_example$s1.a,d_example$s1.b, d_example$s1.c), szero=
cbind(d_example$s0.a, d_example$s0.b, d_example$s0.c), number =
"multiple", type = "robust", var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.474143
##
## $conf.int.quantile.R.s
## [1] 0.413804 0.521525
```

where the only change is for the argument `type = "robust"`. Results show that the estimated PTE is 0.47 (95% CI: 0.42, 0.52).

Notably, these extensions focus on identifying the PTE of a set of potential markers and do not aim to identify *which* surrogate markers are most/least useful in terms of capturing the treatment effect. Indeed, if one were to estimate a high PTE for a set of markers, a reasonable question, from a statistical, regulatory, and clinical perspective, would be to ask which markers may be driving the PTE estimate and similarly, which markers contribute minimally, for example, with a low PTE if considered individually.

Though we do not describe them here, methods exist to estimate the PTE of a high-dimensional surrogate (e.g., changes in gene expression) and a longitudinal surrogate (e.g., repeated measures of fasting plasma glucose). For a high-dimensional surrogate, Zhou et al. (2022) [60] propose a fully parametric approach based on a debiased scaled lasso, with implementation via the `ptehd` function in the `freedbird R` package. Agniel et al. (2023) [61] offer a more flexible doubly robust method to evaluate a high-dimensional surrogate which does not require that the treatment be randomized. The authors develop and implement two versions of the proposed estimator: One based on the Super Learner, [62] which finds an optimal combination of a set of candidate models or learners, and another based on the relaxed lasso [63]. Implementation is available via the `xf_surrogate` function in the `crossurr R` package. For a longitudinal marker, Agniel and Parast (2021) [64] propose multiple approaches to estimate the PTE, one based on a functional linear model, another on a generalized additive model, and a final kernel approach. Implementation is available via the `estimate_surrogate_value` function (with options 'linear', 'gam', or 'kernel') in the `longsurr R` package. Xu and Zeger (2001) [65] propose a parametric latent variable approach for the setting with multiple longitudinal markers and offer an approach to compare different subsets of markers.

4.2 | Measurement Error

When the potential surrogate marker is measured with error, the estimated PTE may be biased. Ignoring measurement error in a surrogate marker can result in misleading conclusions regarding the validity of the surrogate marker, meaning either potentially overvaluing a useless surrogate marker or discarding a valid surrogate marker [66–68]. For example, previous work has shown that bias induced by measurement error in the surrogate marker measurement can result in an estimated PTE of 0.65 (with a 95% CI lower bound less than 0.50), when, in fact, the true PTE is 0.90, implying that the relatively strong surrogate marker would be identified as a poor surrogate. When the parametric models (1) and (2) hold, the attenuation bias due to a mis-measured surrogate can be expressed in a closed form such that the PTE is strictly under-estimated [69]. When the parametric models do not hold and the model-free definition of Wang and Taylor is used instead with non-parametric estimation, attenuation bias is still present but cannot be expressed in a closed form. Parast et al. (2022) [69] provide methods to correct this bias due to measurement error via a straightforward bias correction, in the parametric case, and via the use of simulation extrapolation (SIMEX) more generally. SIMEX estimation is a simulation-based approach to correct measurement error by approximating the bias attributable to measurement error [70, 71]. This method involves two steps: First *more* measurement error is generated and one observes how the error affects the bias of the estimate of interest; then, this association

between error and bias is extrapolated to a setting with no measurement error. In the surrogate setting, SIMEX estimation can be used to understand the magnitude and direction of the bias in the estimation of the PTE that is induced by measurement error and to remove this bias. While SIMEX is a powerful tool, it is important to note it may be unreliable in certain settings such as settings with a small sample size or in which the relationship between the error and the bias is highly non-linear. Methodological details and justification can be found in Parast et al. (2022) [69].

Measurement error correction can be implemented using the `R.s.estimate.me` function within the `Rsurrogate` package as follows. The dataset `d_example_me` within the package is similar to the `d_example` dataset used above with the exception that the surrogate is measured with error. The dataset contains variables (columns) `y1`, `y0`, `s1`, and `s0` which are the primary outcome for the treated group, the primary outcome for the control group, the mis-measured surrogate for the treated group, and the mis-measured surrogate for the control group. Additionally, the dataset contains three replicates each of `s1` and `s0`; these replicates are used to estimate the variance of the measurement error. In general, we need some information about the variance of the measurement error, either it is provided to us externally or we estimate it using, for example, replicates. There are some similarities between the `R.s.estimate.me` and `R.s.estimate` functions. Similar to `R.s.estimate`, the `R.s.estimate.me` function takes in arguments `sone`, `szero`, `yone`, `yzero` and has an `extrapolate` argument which can be set to `TRUE` or `FALSE`, with `TRUE` being the default. Unlike `R.s.estimate`, the `R.s.estimate.me` function has other arguments which we explain and illustrate here. Below, we take a look at the dataset and then estimate the PTE using the `R.s.estimate.me` function.

```
# take a look at the data
data(d_example_me)
names(d_example_me)

##  [1] "y1"      "s1"      "s1_rep1" "s1_rep2" "s1_rep3" "y0"      "s0"
##  [8] "s0_rep1" "s0_rep2" "s0_rep3"

# Estimate the PTE of a surrogate with measurement error using the disattenuated estimator
R.s.estimate.me(yone = d_example_me$y1, yzero = d_example_me$y0, sone =
d_example_me$s1, szero = d_example_me$s0, parametric = TRUE, estimator = "d",
me.variance = 0.5, naive = TRUE, Ronly = FALSE)

## [abbreviated output shown]
## $R.naive
## [1] 0.2161019
##
## $R.naive.CI.normal
## [1] 0.1047344 0.3274694
##
## $R.corrected.dis
## [1] 0.26888
##
## $R.corrected.CI.normal.dis
## [1] 0.1603651 0.3773949
```

When the `parametric` argument is `TRUE`, as above, the PTE is estimated using the Freedman approach; when it is `FALSE` the non-parametric kernel-smoothing approach is used. The argument `estimator` can be either `"d"`, `"q"` or `"n"` where `"d"` stands for disattenuated and is only allowed for the parametric estimation approach (because the attenuation bias can be directly calculated), `"q"` stands for SIMEX correction with quadratic extrapolation, and `"n"` stands for SIMEX correction with non-linear extrapolation. The argument `me.variance` is the variance of the measurement error which must be provided (below, we will use the replicates to estimate it), `naive` is logical and indicates whether the user also wants the naive (uncorrected) estimator to be provided in the output, and `Ronly` is logical and indicates whether the user wants only the R (PTE) estimates only or all estimates including, for example, Δ and Δ_S when non-parametric estimation is used. Depending on the argument specifications, the resulting output can be quite overwhelming though informative. Distinct from `R.s.estimate`, this function does not use resampling for variance estimation (analytic variances from the closed-form derivations are used instead) and thus provides only two types of confidence intervals: One based on a normal approximation and one based on Fieller's theorem. The results from the code above indicate that the corrected PTE estimate is 0.27 with a 95% CI (0.16, 0.38), whereas the naive estimate was 0.22 with a 95% CI (0.10, 0.33). Next, we apply the function using the SIMEX correction with quadratic extrapolation instead of using the disattenuated estimator. Note that here, we need to set a seed for reproducibility because there is randomness in the simulation step of SIMEX; we also set `Ronly` to `TRUE` and `naive` to `FALSE` to reduce output (and since we already have the naive estimator above).

```
# Estimate the PTE of a surrogate with measurement error using the SIMEX estima-
tor with quadratic
# extrapolation
set.seed(5)
R.s.estimate.me(yone = d_example_me$y1, yzero = d_example_me$y0, sone = d_example_me$s1,
szero = d_example_me$s0, parametric = TRUE, estimator = "q", me.variance = 0.5,
naive = FALSE, Ronly = TRUE)

## [abbreviated output shown]
## $R.corrected.q
## [1] 0.2560957
##
## $R.corrected.CI.normal.q
## [1] 0.1997765 0.3124150
```

The estimate corrected via SIMEX is 0.26 with a 95% CI (0.20, 0.31), resulting in a similar estimate as the disattenuated estimator but a tighter confidence interval. Next, we illustrate using the replicates to estimate the measurement error variance, and then providing that variance to the function. In addition, the code below uses the non-parametric estimator and the SIMEX correction with quadratic extrapolation.

```
# create replicates matrix
replicates = rbind(cbind(d_example_me$s1_rep1, d_example_me$s1_rep2, d_example_me$s1_rep3),
cbind(d_example_me$s0_rep1, d_example_me$s0_rep2, d_example_me$s0_rep3))

# estimate measurement error variance
mean.i = apply(replicates, 1, mean)
num.i = apply(replicates, 1, length)
var.u = sum((replicates - mean.i)^2)/sum(num.i)
var.u

## [1] 0.329879

# Estimate the PTE of a surrogate with measurement error using SIMEX estimator with quadratic
# extrapolation
set.seed(5)
R.s.estimate.me(yone = d_example_me$y1, yzero = d_example_me$y0,
sone = d_example_me$s1, szero = d_example_me$s0, parametric = FALSE, estimator = "q",
me.variance = var.u, naive = FALSE, Ronly = TRUE)

## [abbreviated output shown]
## $R.naive
## [1] 0.181793
##
## $R.naive.CI.normal
## [1] 0.1190757 0.2445103
##
## $R.corrected.q
## [1] 0.180115
##
## $R.corrected.CI.normal.q
## [1] 0.09754789 0.26268215
```

The estimated variance, `var.u`, is 0.33. The non-parametric estimate corrected via SIMEX is 0.18 with a 95% CI (0.10, 0.26). In the next section, we address surrogate evaluation when the primary outcome is subject to censoring.

4.3 | Censored Outcomes

In some settings, the primary outcome of interest is a censored time-to-event outcome. For example, the primary outcome may be the time to diabetes diagnosis, time to dementia onset, time to metastasis, or time to death. Of course, many methods have been developed

for censored outcomes, but a censored time-to-event outcome in the surrogate evaluation setting is particularly complicated. Specifically, censoring not only precludes us from observing the primary outcome but also may preclude us from measuring the surrogate marker if the individual is censored before the surrogate is measured. For example, let t_0 be the time that the surrogate marker is measured where $t_0 > 0$; if an individual dies before t_0, we clearly cannot measure the surrogate marker at t_0. Even when the primary outcome is not death, it is often the case that the individual is no longer followed in the study after the primary outcome occurs or the individual starts a different course of treatment after the primary outcome occurs, making the surrogate marker measurement uninformative for our purposes. Uninformative censoring of the surrogate marker alone can be handled using existing methods such as inverse probability weighting, but it is this second complication of the primary outcome occurrence hindering our ability to measure the surrogate that increases complexity. At first glance, one might consider the primary outcome occurring before t_0 to be somewhat of a nuisance. Clearly, it would be inappropriate to evaluate the surrogate marker by simply discarding individuals who experience the primary outcome before t_0. Upon further consideration, recall that the primary outcome *is the outcome* we truly care about. Therefore, Parast et al. (2017) [72] argue that surrogate evaluation in this setting should focus on evaluating the "surrogate information" that is available at t_0 which is a combination of the surrogate marker for those who have not yet experienced the primary outcome before t_0, and the time of the primary outcome for those who have experienced the primary outcome before t_0. While the methodological details of estimating the PTE are necessarily different from the uncensored setting, the idea is similar. The PTE is still defined as the contrast between the treatment effect and the residual treatment effect, but these are now defined in terms of the difference in survival (or cumulative incidence) at a pre-specified time t. For example, the treatment effect is defined as

$$\Delta(t) = P(Y^{(1)} > t) - P(Y^{(0)} > t)$$

The residual treatment effect requires more notation and is omitted here, but notably is a function of both t and t_0, denoted as $\Delta_S(t, t_0)$. The PTE is then defined as $R_S(t, t_0) = 1 - \Delta_S(t, t_0)/\Delta(t)$ and estimation can be implemented fully non-parametrically using the non-parametric kernel Nelson-Aalen estimator, [73, 74] that is, the survival version of the Nadaraya-Watson conditional mean estimator. Methodological details and justification are provided in Parast et al. (2017) [72].

Estimation can be implemented using the `R.s.surv.estimate` function within the `Rsurrogate` package. The dataset `d_example_surv` within the package is similar to the `d_example` dataset used above with the exception that the primary outcome is a censored time-to-event outcome. Specifically, the dataset contains variables (columns) `s1` and `s0` which contain the surrogate marker measured at time t_0 for the treated and control group, respectively. The variables `x1` and `x0` contain the observed event time (the minimum of the true event time and the censoring time), and `delta1` and `delta0` contain the event indicators (1 if the corresponding observed event time is an event, 0 if it is the censoring time) for the treated and control group, respectively. The variables `z1` and `z0` contain a baseline covariate for the treated and control group, respectively. Below, we take a look at the dataset and estimate the PTE using $t = 1$ and $t_0 = 0.5$ where the argument `landmark` denotes the t_0 time; we set the `var` and `conf.int` arguments to `TRUE` indicating that we would like the variance and confidence interval estimates which are estimated using resampling.

```
# take a look at the data
data(d_example_surv)
names(d_example_surv)

## [1] "s1"     "x1"     "delta1" "s0"     "x0"     "delta0" "z1"     "z0"

# Estimate PTE for the censored outcome
set.seed(4)
R.s.surv.estimate(xone = d_example_surv$x1, xzero = d_example_surv$x0, deltaone
= d_example_surv$delta1, deltazero = d_example_surv$delta0, sone = d_example_surv$s1,
szero = d_example_surv$s0, t = 2, landmark = 1, var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.7974101
##
## $R.s.var
## [1] 0.02157223
##
## $conf.int.quantile.R.s
## [1] 0.5371597 1.1020215
```

The function returns estimates for $\Delta(t)$, $\Delta_S(t, t_0)$, and $R_S(t, t_0)$. The estimated PTE here is $\widehat{R}_S(t, t_0) = 0.80$ with 95% CI (0.54, 1.10). Importantly, this function only implements the non-parametric estimator; there are no model-based options. Because this method defines the surrogate information as the combination of the surrogate and primary outcome information up to t_0, it would be reasonable

to ask how much of the estimated PTE is attributable to the surrogate marker itself, rather than the primary outcome. This is referred to as the "incremental value" of S in Parast et al. (2017) [72]. This can be obtained from the function by setting the argument `incremental.value = TRUE` which also calculates the PTE of the primary outcome information only up to t_0 and then takes the difference, along with providing corresponding confidence intervals.

```
# with incremental value
set.seed(4)
R.s.surv.estimate(xone = d_example_surv$x1, xzero = d_example_surv$x0, deltaone =
d_example_surv$delta1, deltazero = d_example_surv$delta0, sone = d_example_surv$s1,
szero = d_example_surv$s0, t = 2, landmark = 1, var = TRUE, conf.int = TRUE,
incremental.value = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.7974101
##
## $conf.int.quantile.R.s
## [1] 0.5371597 1.1020215
##
## $R.t
## [1] 0.5314994
##
## $incremental.value
## [1] 0.2659107
##
## $conf.int.quantile.iv
## [1] 0.0842032 0.4562116
```

The estimated PTE of the primary outcome up to t_0 only is 0.53 such that the incremental value of the surrogate is 0.27 with 95% CI (0.08, 0.46), implying that the surrogate marker does indeed provide substantial incremental value in terms of explaining the treatment effect on $\Delta(t)$.

The package also offers the option to potentially improve efficiency via augmentation by taking advantage of baseline covariates which, by randomization, are not associated with the treatment but may be associated with the primary outcome. This can be implemented with the `Aug.R.s.surv.estimate` function where the additional arguments are `basis.delta.one` and `basis.delta.zero` which are the basis transformations used for augmentation. In the simplest case, these arguments can simply be the baseline covariates. We implement one example of augmentation below using the `z1` and `z0` variables in `d_example_surv`.

```
# with augmentation
set.seed(4)
Aug.R.s.surv.estimate(xone = d_example_surv$x1, xzero = d_example_surv$x0, deltaone
= d_example_surv$delta1, deltazero = d_example_surv$delta0, sone = d_example_surv$s1,
szero = d_example_surv$s0, t = 2, landmark = 1, basis.delta.one = d_example_surv$z1,
basis.delta.zero = d_example_surv$z0)

## [abbreviated output shown]
## $aug.R.s
##           [,1]
## [1,] 0.7974804
##
## $aug.R.s.var
##             [,1]
## [1,] 0.02154302
##
## $conf.int.quantile.aug.R.s
## [1] 0.5374258 1.1017556
```

In this dataset, augmentation offers little-to-no advantage with an estimated variance for $\widehat{R}_S$ of 0.02154 with augmentation and 0.02157 without augmentation. Similar to the above, the function also has the option to set `incremental.value = TRUE`.

These functions evaluate a single surrogate when the primary outcome is a censored time-to-event outcome. If one wishes to evaluate multiple surrogates, as in Section 4.1, this can be implemented with the `R.multiple.surv` function, using the `d_example_multiple` dataset which is similar to `d_example_surv` except that `s1` and `s0` contain four surrogates and the dataset is structured as a list. Below, we take a look at the data and implement the function using $t = 1$ and $t_0 = 0.5$.

```
# take a look at the data
data(d_example_multiple)
names(d_example_multiple)

## [1] "s1"     "x1"     "delta1" "s0"     "x0"     "delta0"

# Estimate PTE of multiple surrogates with a censored outcome
set.seed(4)
R.multiple.surv(xone = d_example_multiple$x1, xzero = d_example_multiple$x0, deltaone =
d_example_multiple$delta1, deltazero = d_example_multiple$delta0, sone =
as.matrix(d_example_multiple$s1), szero = as.matrix(d_example_multiple$s0), t = 1, land-
mark = 0.5,
var = TRUE, conf.int = TRUE)

## [abbreviated output shown]
## $R.s
## [1] 0.6445202
##
## $conf.int.quantile.R.s
## [1] 0.4547815 0.8305558
```

The estimated PTE for this set of surrogates is 0.64 with 95% CI (0.45, 0.83). The function offers six different types of estimators which can be set with the argument `type`: A two-stage robust estimator, a weighted two-stage robust estimator, a double-robust estimator, a two-stage model-based estimator, a weighted estimator, and a double-robust model-based estimator, where the default is the two-stage robust estimator (used in the code above). Details about these estimators and justification are available in Parast et al. (2021) [75].

In the case that, similar to the primary outcome, the surrogate marker itself is a censored time-to-event outcome, the PTE can be evaluated using the `R.q.event` function in the `SurrogateOutcome` package. Details are omitted here but can be found in Parast et al. (2020) [76]. Lastly, influence function-based methods to estimate the PTE of a longitudinal surrogate for a censored outcome, with plug-in and targeted minimum loss-based estimation options, have been recently proposed in Parast and Agniel (2024) [77].

Of course, many alternative methods for the censored outcome setting exist beyond those described above, though they are generally model-based. For example, Lin et al. (1997) [31] examined extending the Freedman approach to the censored outcome setting via Cox proportional hazards models, but showed that it is impossible for both of the specified survival models to hold simultaneously. When both the primary outcome and surrogate are censored outcomes, Ghosh (2008, 2009) [38, 78] proposed estimation and inference procedures for the PTE using copula and accelerated failure time models and demonstrated desirable finite sample performance when the AFT model holds. In the principal stratification framework, Conlon et al. (2017) [47] proposed using a Gaussian copula model with a Bayesian estimation approach, while Gabriel and Gilbert (2014) [79] and Gabriel et al. (2015) [80] proposed flexible procedures using Weibull time-to-event models for the primary outcome. More recently, Roberts et al. (2024) [46] proposed a multi-state illness-death model framework for a censored outcome and surrogate, and Roberts et al. (2023) [44] proposed a mixed-modeling approach for a setting with a longitudinal primary outcome resulting in a surrogate-dependent treatment efficacy curve that allows one to validate the surrogate at different time points.

4.4 | Heterogeneity in the Utility of a Surrogate

Similar to the concept of treatment effect heterogeneity, heterogeneity in the utility of a surrogate is when the surrogate is a valid replacement of the primary outcome for some subgroup of the population, but not for others. This is especially problematic if the surrogate is to be used to replace the primary outcome in a future trial that may consist of a different mix of patients [81]. Recall that such replacement is the ultimate goal in identifying surrogate markers and certainly, one does typically expect different trials to involve different types of patients. For ease of notation, let W denote a measured baseline covariate, such as age. Here, we describe a definition and estimation approach for the PTE *as a function of* $W = w$, proposed in Parast et al. (2023) [82]. Let the treatment effect as a function of w be

$$\Delta(w) = E(Y^{(1)}|W = w) - E(Y^{(0)}|W = w)$$

and the residual treatment effect as a function of w be

$$\Delta_S(w) = \int E(Y^{(1)} - Y^{(0)} | S^{(1)} = S^{(0)} = s, W = w) dF_0(s|w)$$
$$= \int \mu_1(s,w) dF_{S^{(0)}}(s|w) - \int \mu_0(s,w) dF_{S^{(0)}}(s|w)$$

where $\mu_g(s,w) = E(Y^{(g)}|S^{(g)} = s, W = w)$, $F_{S^{(0)}}(\cdot|w)$ is the cumulative distribution function of $S^{(0)}$ given $W = w$, and the second equality follows from a needed identifiability assumption that $Y^{(1)} \perp S^{(0)} | S^{(1)}, W$ and $Y^{(0)} \perp S^{(1)} | S^{(0)}, W$. Thus, the PTE as a function of $W = w$ is defined as $R_S(w) = 1 - \Delta_S(w)/\Delta(w)$. When W is univariate, $R_S(w)$ can be estimated fully non-parametrically. Specifically, with a continuous W, $\Delta(w)$ can be estimated as

$$\hat{\Delta}(w) = \hat{\mu}_1(w) - \hat{\mu}_0(w)$$

where

$$\hat{\mu}_g(w) = \frac{\sum_{i=1}^{n_g} K_{h_g}(W_{gi} - w) Y_{gi}}{\sum_{i=1}^{n_g} K_{h_g}(W_{gi} - w)}, g = 0, 1$$

where h_1 and h_0 are bandwidths, which may be data dependent. The quantity $\Delta_S(w)$ can be estimated using two-dimensional smoothing as

$$\hat{\Delta}_S(w) = \hat{\mu}_{10}(w) - \hat{\mu}_0(w)$$

where $\hat{\mu}_{10}(w) = \int \hat{\mu}_1(s,w) d\hat{F}_{S^{(0)}}(s|w)$,

$$\hat{F}_{S^{(0)}}(s|w) = \frac{\sum_{i=1}^{n_0} K_{h_2}(W_{0i} - w) I(S_{0i} \le s)}{\sum_{i=1}^{n_0} K_{h_2}(W_{0i} - w)},$$

$$\text{and} \quad \hat{\mu}_1(s,w) = \frac{\sum_{i=1}^{n_1} K_{h_3}(S_{1i} - s) K_{h_4}(W_{1i} - w) Y_{1i}}{\sum_{i=1}^{n_1} K_{h_3}(S_{1i} - s) K_{h_4}(W_{1i} - w)}$$

are non-parametric smoothed estimators of the conditional cumulative distribution of $S^{(0)}$ given $W = w$, and the conditional expectation of $Y^{(1)}$ given $(S^{(1)}, W) = (s, w)$, respectively. The bandwidths $h_k, k = 0, \ldots, 4$ may be calculated as described in Section 3. Finally, $R_S(w)$ can be estimated as $\hat{R}_S(w) = 1 - \hat{\Delta}_S(w)/\hat{\Delta}(w)$. Parallel estimation methods for the case, when W is discrete, are also available [82].

In addition, one can formally test for heterogeneity. Specifically, the hypotheses of interest would be:

$$H_0 : R_S(\cdot) \text{ is constant within } [w_a, w_b], i.e., \exists \tau$$
$$\text{s.t. } \forall w \in [w_a, w_b], R_S(w) = \tau,$$
$$H_A : R_S(\cdot) \text{ is not constant within } [w_a, w_b], i.e., \forall \tau, \exists w \in [w_a, w_b]$$
$$\text{s.t. } R_S(w) \neq \tau$$

That is, the null hypothesis is that there is no heterogeneity, that is, $R_S(w)$ is the samefor all w. Parast et al. (2023) [82] offer two testing procedures, an omnibus test based on a supreme-type test statistic that tests across a broad range of alternatives, and a trend test that can provide more power if one is willing to assume that $R_S(w)$ is monotone in w. Both tests as well as the estimation of $R_S(w)$ are implemented in the `hetsurr` package which we illustrate here. The example dataset in this package is `example.data` with variables `y1`, `y0`, `s1`, `s0`, `w1`, `w0`, which are the primary outcomes, surrogate markers, and baseline covariate W for the treatment and control groups, respectively. The `hetsurr.fun` function estimates $R_S(w)$ on a grid of w; if this grid is not specified by the user, the default is to use 50 equally spaced points between the 10th and 90th percentile of the union of `w1` and `w0` values (if continuous) or the set of unique categories (if discrete). At a minimum, this function expects the primary outcome, surrogate marker, and W values for both groups. The function, illustrated below, returns estimates of $\Delta(w)$, `delta.w`, estimates of $\Delta_S(w)$, `delta.w.s`, and estimates of $R_S(w)$, `R.w.s`, for each value on the w grid. When `var=TRUE` as we have below, the function also returns variance estimates for each quantity and associated 95% CIs; for `R.w.s`, pointwise confidence intervals and a confidence band are provided. Because estimation is based on two-dimensional smoothing and resampling is used for variance estimation, the computation time may be long. The output can be overwhelming, thus, we can use the `hetsurr.plot` function to visualize the results.

```
#install the package from CRAN
install.packages("hetsurr")

#load the R package
```

```
library(hetsurr)

#take a look at the data
data(example.data)
names(example.data)

## [1] "y1" "y0" "s1" "s0" "w1" "w0"

# Examine heterogeneity in the utility of the surrogate
set.seed(5)
het.ob = hetsurr.fun(y1 = example.data$y1, y0 = example.data$y0, s1 = example.data$s1,
s0 = example.data$s0, w1 = example.data$w1, w0 = example.data$w0, var = TRUE)

# take a look at values returned in het.ob
names(het.ob)

##  [1] "w.values"             "delta.w"              "delta.w.s"
##  [4] "R.w.s"                "se.delta.w"           "se.delta.w.s"
##  [7] "se.R.w.s"             "conf.delta.w.lower"   "conf.delta.w.upper"
## [10] "conf.delta.w.s.lower" "conf.delta.w.s.upper" "conf.R.w.s.lower"
## [13] "conf.R.w.s.upper"     "band.R.w.s.lower"     "band.R.w.s.upper"

# Plot estimates to visualize heterogeneity
hetsurr.plot(het.ob)
```

The plot is quite large and if you are using RStudio, you may receive a message that the figure margins are too large. The `png` function below directly saves the plot to your working directory (or directly in with your R markdown file if using the R markdown file), where the width and height are specified in inches but can be easily adjusted, and the resolution is set to 300 dots per inch. This plot is shown in Figure 2.

```
 # To save the high-resolution plot directly to a file called plot.png
png("plot.png", width = 8, height = 11, units = "in", res = 300)
hetsurr.plot(het.ob)
dev.off()
```

The formal test of heterogeneity can be implemented with the `test = TRUE` argument as follows:

```
#asking for tests for heterogeneity
set.seed(5)
het.ob = hetsurr.fun(y1 = example.data$y1, y0 = example.data$y0, s1 = example.data$s1,
s0 = example.data$s0, w1 = example.data$w1, w0 = example.data$w0, test = TRUE)

# values returned in het.ob, now with testing results
names(het.ob)
het.ob$omnibus.p.value

##  [1] "w.values"               "delta.w"                "delta.w.s"
##  [4] "R.w.s"                  "se.delta.w"             "se.delta.w.s"
##  [7] "se.R.w.s"               "conf.delta.w.lower"     "conf.delta.w.upper"
## [10] "conf.delta.w.s.lower"   "conf.delta.w.s.upper"   "conf.R.w.s.lower"
## [13] "conf.R.w.s.upper"       "band.R.w.s.lower"       "band.R.w.s.upper"
## [16] "omnibus.test.statistic" "omnibus.p.value"        "trend.test.statistic"
## [19] "trend.p.value"

## [1] 0.018
```

Testing results (test statistic and p-value) for both the omnibus test and trend test are provided. For example, here the omnibus test results in a p-value of 0.018. When W is discrete, estimation and testing can be similarly implemented using the same function by using `type = "discrete"`.

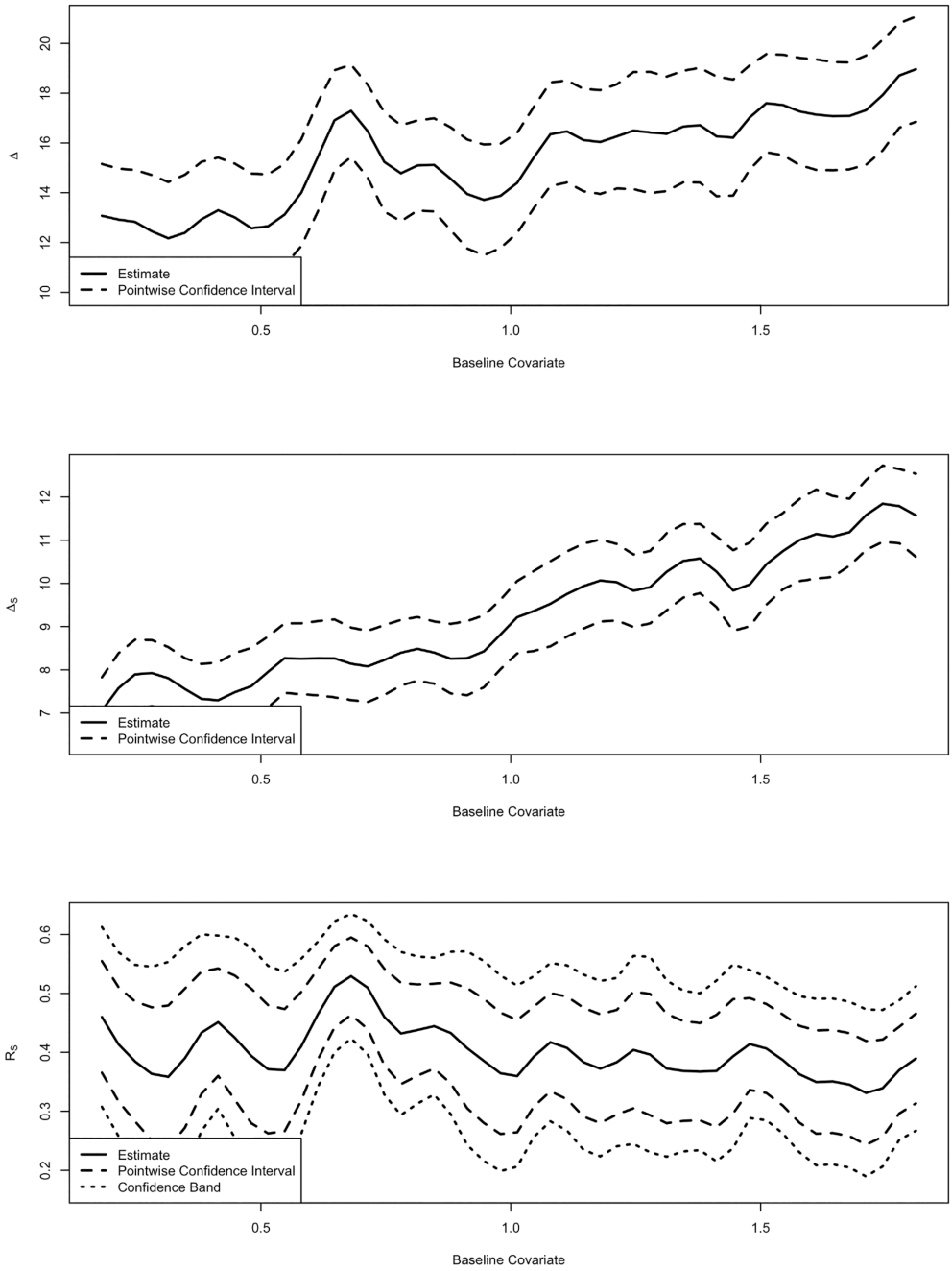

FIGURE 2 | Visualization of heterogeneity results using the `hetsurr` package showing estimates of $\Delta(w)$ (top panel), $\Delta_S(w)$ (middle panel) and $R_S(w)$ (bottom panel) with pointwise confidence intervals (for all; dashed lines) and bands (for $R_S(w)$ only; dotted lines).

While heterogeneity in surrogate utility has been less well-studied than surrogacy in general, some alternative methods do exist, though again, they tend to be model-based. For example, Roberts et al. (2021) [83] propose an approach to define and estimate the causal effect predictiveness curve within patient subgroups via a multivariate normal model. Within a meta-analytic framework, Elliott et al. (2015) [84] provide methods to estimate the relative effect and adjusted association stratified by a baseline covariates via a hierarchical bivariate normal model. Also within a multiple studies setting but with a censored primary outcome, Sachs et al. (2024) [85] utilize a novel hierarchical Bayesian semiparametric model for surrogacy evaluation and to identify potential clusters with differential surrogacy. In a single study setting with a censored outcome, Parast et al. (2024) [86] offer a non-parametric extension of the approach described above to examine and test for heterogeneity, with implementation available in the `hetsurrSurv` package. When W contains multiple covariates, a fully non-parametric procedure is not feasible. In this case, Knowlton et al. (2024) [87] investigate and compare a fully parametric and a semiparametric single-index approach with procedures to identify regions of strong surrogacy, with an implementation available in the `cohetsurr` package.

4.5 | Surrogate Evaluation From a Prediction Perspective

As described previously, our ultimate goal in evaluating a surrogate is to be able to use it in a future study to replace the primary outcome to make inferences on the treatment effect. Thus, a reasonable perspective is to think of surrogate evaluation as a prediction problem such that, in the future, we would like to predict the treatment effect on Y using the treatment effect on S. With this prediction perspective, a valid surrogate should be one that can provide a good prediction. Multiple methods have been developed with this perspective. Price et al. (2018) [88] propose the idea of finding the optimal function of a potential surrogate as that which satisfies the Prentice definition (described in Section 1) of a valid surrogate and optimally predicts the final outcome, with super-learner and targeted super-learner based estimators. Athey et al. (2019) [35] proposes the construction of a surrogate index such that when the Prentice criterion holds, the treatment effect on the surrogate index equals the treatment effect on the outcome, with extension to a non-randomized setting. Wang et al. (2020) [89] propose a method to identify the optimal transformation of a potential surrogate such that the treatment effect on the transformed surrogate closely approximates the treatment effect on the primary outcome. Specifically, the optimal transformation, denoted as $g_{opt}(\cdot)$, minimizes the mean squared error loss function:

$$\mathcal{L}(g_{opt}) = E[(Y^{(1)} - Y^{(0)}) - \{g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})\}]^2$$

and then the treatment effect on $g_{opt}(S)$, $\Delta_{g_{opt}}(S) = E\{g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})\}$ is used to approximate the treatment effect on Y, Δ. They then define the PTE of $g_{opt}(S)$ as simply $\Delta_{g_{opt}}(S)/\Delta$ and show that this definition parallels R_S above, but with an optimally chosen reference distribution. Estimation of $g_{opt}(S)$, $\Delta_{g_{opt}}(S)$, and the resulting PTE can be implemented fully non-parametrically via the `OptimalSurrogate` package, which can accommodate both a continuous or discrete surrogate.

Within the `OptimalSurrogate` package, the primary estimation function is `pte_cont` and we illustrate its use below using the dataset `marker_cont` within the package. This dataset contains three variables: `sob` which is the surrogate marker, `yob` which is the primary outcome, and `aob` which is the treatment indicator where 1 indicates treatment and 0 indicates control. Recall that we have assumed in this paper that higher values of Y and S are "better"; in this dataset, *lower* values of Y are "better" and thus, to be consistent with the rest of the paper, we multiply `yob` by −1. The function requires, at a minimum, the arguments `sob`, `yob`, and `aob` which map to the descriptions above. Variance estimation and confidence intervals are resampling-based and can be requested using the `var` and `conf.int` logical arguments. The function returns a table of estimates (`Estimates`) of Δ, $\Delta_{g_{opt}}(S)$, and two versions of the PTE, which we discuss below, with standard errors and 95% confidence intervals. The function also returns a matrix (`Transformed.S`) which has the same number of rows as `sob` and contains the original values of `sob` in the first column, the estimated optimal transformation of `sob` in the second column, and the standard error, and lower and upper bound of the 95% confidence interval for the transformation in the third, fourth, and fifth columns, respectively. Below, we implement the function and take a look at the estimates.

```
#install the package from CRAN
install.packages("OptimalSurrogate")

# load the package
library(OptimalSurrogate)

# take a look at the data
data(marker_cont)
names(marker_cont)

## [1] "sob" "yob" "aob"

#flip so that higher values of Y are better
marker_cont_tutorial = marker_cont
marker_cont_tutorial$yob = -1*marker_cont$yob

# Estimate the optimal transformation of S and the PTE of the optimal transformation
set.seed(5)
out = pte_cont(sob = marker_cont_tutorial$sob, yob = marker_cont_tutorial$yob, aob =
marker_cont_tutorial$aob,  var = TRUE, conf.int = TRUE)

#take a look at estimates
out$Estimates

##                est         se     lower     upper
## delta    0.2412751 0.03042336 0.1816453 0.3009049
## delta.gs 0.1597523 0.02030530 0.1199539 0.1995507
## pte1     0.6621168 0.05654075 0.5512969 0.7729367
## pte2     0.4927234 0.02711679 0.4395745 0.5458723
```

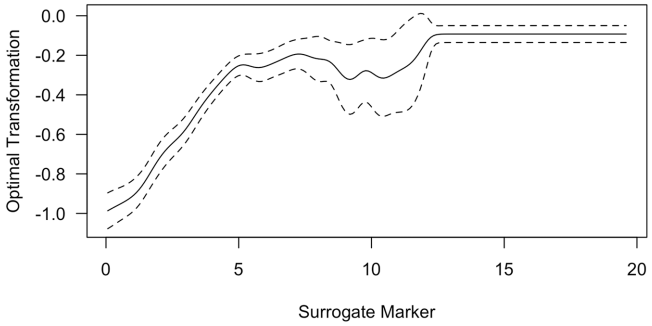

FIGURE 3 | Visualization of the optimal transformation of the surrogate (solid line) using results from the `OptimalSurrogate` package with pointwise confidence intervals (dashed lines).

Results show that the estimated Δ is 0.24 with 95% CI (0.18, 0.30), the estimated $\Delta_{g_{opt}}(S)$ is 0.16 with 95% CI (0.12, 0.20), and the estimated PTE (`pte1`) defined above is 0.66 (0.55, 0.77). The estimates contained in `pte2` are for a slightly different definition of PTE which is based on the percentage of *variation* in $Y^{(1)} - Y^{(0)}$ which is explained by the variation in $g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})$ with estimation involving calculating the mean square error of various terms. Next, we use the estimates contained in `Transformed.S` to visualize the transformed surrogate, shown in Figure 3.

```
# visualize the transformed surrogate
x = as.numeric(rownames(out$Transformed.S))
plot(x, out$Transformed.S[, "est"], ylim = range(out$Transformed.S[, -2]), type = "l",
las = 1, xlab = "Surrogate Marker", ylab = "Optimal Transformation")
lines(x, out$Transformed.S[, "lower"], lty = 2)
lines(x, out$Transformed.S[, "upper"], lty = 2)
```

This approach can be quite advantageous compared to the untransformed approach in Section 3 since it is more general, identifying an optimal transformation of the surrogate as well as using a reference distribution for the PTE that corresponds to this optimal transformation, though may be considered more difficult to interpret in practice. This optimal transformation approach has been extended to a censored primary outcome setting [90] with implementation available in the `OSsurivival` package, and the multiple surrogates setting [91] with implementation available in the `CMFsurrogate` package.

4.6 | Surrogate Paradox

We return now to the discussion of arrhythmia drugs from Section 1. As an admittedly simplified description of a very complicated situation, anti-arrhythmia drugs were approved by the FDA based on evidence that they had an effect on a surrogate when, in fact, the drugs killed patients. While researchers can propose countless novel statistical methods to evaluate a surrogate, arguably the most important point is to ensure that this does not happen again. This situation is referred to as the surrogate paradox, where there is a positive association between the surrogate and the outcome, a positive treatment effect on the surrogate, but then a negative effect on the outcome [34, 92, 93]. In general, the methods described above must impose certain assumptions to ensure "protection from" the surrogate paradox. For example, the PTE framework generally imposes these three sufficient but not necessary conditions:

A1. $P(S^{(1)} > s) \geq P(S^{(0)} > s)$ for all s,

A2. $\mu_1(s)$ and $\mu_0(s)$ are monotone increasing in s; recall that $\mu_g(s) = E(Y^{(g)}|S^{(g)} = s)$,

A3. $\mu_1(s) \geq \mu_0(s)$ for all s.

It can be shown that with these conditions, the surrogate paradox cannot occur and $R_W \in [0, 1]$ [94]. These conditions are testable in the study being used to evaluate the surrogate and formal non-parametric tests of each condition in a single study have been proposed in Hsiao et al. (2024) [94] with implementation available in the `SurrogateParadoxTest` package which we illustrate below. Specifically, we use the `test_assumptions` function to test (A1)–(A3) in the `d_example` dataset used in Section 3. The function refers to condition (A1), (A2), and (A3) as stochastic dominance, monotonicity, and non-negative residual treatment effect, respectively, and proposes a formal test for each. The minimum required arguments are `s0`, `s1`, `y0`, and `y1`, which are the surrogates in the control and treatment groups and the outcomes in the control and treatment groups, respectively. The argument `monotonicity_bootstrap_n` indicates the number of bootstrap replications to be used for the monotonicity test, that is, condition (A2), and the argument `nnr_bootstrap_n` indicates the number of bootstrap replications to be used for the non-negative residual treatment effect test, that is, condition (A3). Both of these tests are computationally intensive; thus, for illustration in this tutorial we have set the

number of bootstrap replications for both to be small but in practice, they should be set to a more typical bootstrap replication number such as 200. The `all_results=FALSE` argument limits the return output to a simple table showing the results for each test which we include below.

```
#install the package from CRAN
install.packages("SurrogateParadoxTest")

# load package
library(SurrogateParadoxTest)

# test assumptions
set.seed(5)
test_assumptions(s0 = d_example$s0.a, y0 = d_example$y0, s1 = d_example$s1.a,
y1 = d_example$y1, all_results = FALSE, monotonicity_bootstrap_n = 5, nnr_bootstrap_n = 5)

#results
  Assumption                                Result
1 "Stochastic dominance assumption"         "Holds"
2 "Monotonicity assumption (control)"       "Holds"
3 "Monotonicity assumption (treatment)"     "Holds"
4 "Non-negative residual treatment effect" "Holds"
```

The output states the result of each test in terms of whether the assumption/condition holds. In this example, the results of the tests are that the stochastic dominance assumption, condition (A1), holds; the monotonicity assumption in both the control and treatment group, condition (A2), holds; and the non-negative residual treatment effect, condition (A3), holds.

To summarize, these tests indicate whether (A1)–(A3) hold in the study being used to evaluate the surrogate, which we will refer to as the current study. However, we need these assumptions to hold in the study where the surrogate *is to be used to test* for a treatment effect. Of course, if this future study parallels the current study exactly, then the results of these tests are exactly what is needed. While there may be some settings where this is true, generally such a transportability assumption would be extremely strong. Work is ongoing to address testing these assumptions in a future study. Existing work focused on the future study includes Elliott et al. (2015) [84] which proposes an approach to directly estimate the probability of the surrogate paradox occurring in the future study given (a) multiple prior studies and (b) the treatment effect on the surrogate marker in the future study.

5 | Discussion

In this tutorial, we have described available frameworks for evaluating a surrogate marker while focusing on the practical implementation of the PTE framework and covering uncensored and censored outcomes, parametric and non-parametric estimation, evaluating multiple surrogates, heterogeneity in the utility of the surrogate marker, surrogate validation from a prediction perspective, and the surrogate paradox.

Importantly, there is a close link between surrogate evaluation and mediation analysis [29, 34, 61, 95, 96]. In mediation, the aim is to determine whether the effect of a treatment on a primary outcome is mediated by a particular intermediary variable, such that the intermediary variable can explain how or why the treatment affects the primary outcome. By definition, the question of mediation is a causal question and a true mediator must be on the causal pathway between the treatment and the primary outcome. At a high-level, the underlying math is similar in surrogate evaluation and mediation analysis and in a randomized setting, methods developed for mediation can often be used for surrogate evaluation. Broadly speaking, both aim to determine whether a third variable accounts for the effect of a treatment on an outcome, often resulting in a measure that quantifies the proportion (or amount) of the treatment effect explained by this third variable. However, at a deeper level, they are very different. A surrogate does *not* need to be on the causal pathway between the treatment and the outcome; there is no inherent requirement that the surrogate explains the mechanism behind the treatment's efficacy. In an extreme case, a variable can be superficially correlated with something that is on the causal pathway and it would potentially be a perfect surrogate but not at all a mediator. In addition, typically, the actions that are taken after surrogate evaluation vs. after a mediation analysis are very different. In surrogate evaluation, a variable that is identified as a surrogate is then considered as a candidate outcome in a future trial, that is, the primary outcome may be replaced by the surrogate in terms of evaluating the effect of a treatment. In mediation analysis, it is rarely the case that an identified mediator would be argued to replace the outcome in a future study; that is, the primary objective in mediation analysis is *not* to replace the outcome, but to explain the mechanisms behind the treatment effect on the outcome. For example, the mediation analysis in Croce et al. (2024) [97] examined the extent to which socioeconomic factors such as food insecurity explain the effect of race on asthma prevalence in a pediatric population. There is no intent here to argue for replacing asthma prevalence with food insecurity in future studies. Instead, the aim is to examine whether, how,

and by how much socioeconomic factors explain disparities in asthma prevalence. This example highlights another typical difference: Surrogate evaluation is generally conducted in randomized studies whereas mediation analysis is often, though of course not always, conducted in observational studies. Of course, the complications that come with observational studies are immense; the inability to assume that treatment is randomized typically makes the straightforward application of surrogacy methods to mediation inappropriate. For example, even estimation of Δ from Section 3 would need to be different if treatment was not randomized and instead would need to involve methods that adjust for selection bias, such as propensity score methods. We contend that mediation analysis is considerably more challenging than surrogate evaluation; however, the decisions based on surrogate evaluation results carry far greater significance than those arising from mediation analysis. A mediation analysis that concludes that an intermediary variable explains 70% of the treatment effect when it really explains only 40% would lead to incorrect attribution of the mechanism behind the treatment effect. In contrast, a surrogate evaluation that concludes that a potential surrogate explains 70% of the treatment effect when it really explains only 40% would lead to a future trial using a poor surrogate to replace the outcome and possibly result in that trial making dramatically incorrect conclusions about the treatment effect.

There are many open problems in this research area. First and foremost is the question of which surrogate evaluation method to use in practice. As is evident in this tutorial, our bias leans towards the PTE framework. While other frameworks offer many advantages, the intuition and appeal of the PTE quantity make it particularly attractive, as evidenced by its continued use in practice—including the parametric Freedman approach, which has repeatedly been shown to be inappropriate [98–102]. Of course, widespread adoption does not inherently validate a method's correctness. However, we argue that a more practical path forward is to develop improved methods for estimating the PTE rather than introducing entirely different frameworks that are less likely to gain traction among researchers actively evaluating surrogates. Ideally, all existing frameworks should be applied to a potential surrogate, and only if there is general agreement across frameworks that the variable is a valid surrogate should it be considered for use in future trials. (Of course, if only a single study is available, formally, the meta-analytic framework is not applicable.) Second, there is little research addressing the subject of missing data in surrogate evaluation, for example, missingness in the surrogate, primary outcome, or both. In general, the PTE-specific packages illustrated above do not allow for missing data. The limited methods that are available focus on the principal stratification framework and/or involve strong parametric assumptions [103–105]. Future work to appropriately accommodate missingness would be a valuable contribution to this research area. Lastly, as the ultimate goal in evaluating a surrogate is to then use a valid surrogate to replace the outcome in a future study, an important open problem is how to and when is it appropriate to test for a treatment effect in the future study using the surrogate marker. While some statistical methods have been proposed for the testing component, [35, 81, 90, 106–108] limited work has been done on the issue transportability of surrogate knowledge from one study to another [109, 110]. That is, even if a surrogate is shown to be valid in a prior study, what justification do we have to use that surrogate in a future study if that future study may look different with respect to the patient population, treatment, or something else? This issue in the surrogate setting is directly related to transfer learning and domain adaptation, [111–113] research areas that have seen an immense amount of progress in recent years. Future work on the transportability of surrogate information is warranted.

Evaluating surrogate markers can indeed be a daunting and discouraging task [15, 34, 114]. It often requires relying on multiple unverifiable assumptions, and errors in determining surrogate validity can and do have serious consequences. In an ideal world, one might argue that surrogate evaluation should be avoided altogether and that only primary outcomes should be used for formal drug approval. However, this approach is simply impractical given the immense pressure to develop, test, and approve new effective treatments. Striking a balance between scientific rigor and practicality is essential, and continued efforts to refine surrogate evaluation methods are crucial for ensuring both the safety and efficiency of drug development.

Acknowledgements

Support for this research was provided by the National Institutes of Health grant R01DK118354.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

The author has nothing to report.

References

1. FDA, "Fast Track, Breakthrough Therapy, Accelerated Approval," *Priority Review: Accelerated Approval* (2024), https://www.fda.gov/patients/learn-about-drug-and-device-approvals/fast-track-breakthrough-therapy-accelerated-approval-priority-review.

2. Lilly, "FDA Approves Jardiance (Empagliflozin) Tablets for Adults With Type 2 Diabetes," 2014, https://investor.lilly.com/news-releases/news-release-details/fda-approves-jardiancer-empagliflozin-tablets-adults-type-2.

3. FDA, "FDA Approves Novel, Dual-Targeted Treatment for Type 2 Diabetes," 2016, https://www.fda.gov/news-events/press-announcements/fda-approves-jardiance-reduce-cardiovascular-death-adults-type-2-diabetes.

4. B. Dunn, P. Stein, and P. Cavazzoni, "Approval of Aducanumab for Alzheimer Disease—The FDA's Perspective," *JAMA Internal Medicine* 181, no. 10 (2021): 1276–1278.

5. J. Sevigny, P. Chiao, T. Bussière, et al., "The Antibody Aducanumab Reduces Aβ Plaques in Alzheimer's Disease," *Nature* 537, no. 7618 (2016): 50–56.

6. H. Fillit and A. Green, "Aducanumab and the FDA—Where Are We Now?," *Nature Reviews Neurology* 17, no. 3 (2021): 129–130.

7. G. C. Alexander, D. S. Knopman, S. S. Emerson, et al., "Revisiting FDA Approval of Aducanumab," *New England Journal of Medicine* 385, no. 9 (2021): 769–771.

8. J. Steenhuysen and D. Beasley, "U.S. Approval of Biogen Alzheimer's Drug Sends Shares Soaring, Hailed as Big Day for Patients," 2021, https://www.reuters.com/business/healthcare-pharmaceuticals/us-fda-set-rule-controversial-biogen-alzheimers-drug-2021-06-07/.

9. A. Frankel, "Biogen and Its Investors Both Want US Appeals Court to Clarify Class Certification," 2022, https://www.reuters.com/legal/government/column-biogen-its-investors-both-want-us-appeals-court-clarify-class-2024-10-02/.

10. P. Belluck, "Medicare Officially Limits Coverage of Aduhelm to Patients in Clinical Trials," 2022, https://www.nytimes.com/2022/04/07/health/aduhelm-medicare-alzheimers.html.

11. R. Sachs, *Understanding Medicare's Aduhelm Coverage Decision* (Health Affairs Forefront, 2022).

12. Biogen, "Biogen to Realign Resources for Alzheimer's Disease Franchise," 2024, https://investors.biogen.com/news-releases/news-release-details/biogen-realign-resources-alzheimers-disease-franchise.

13. P. R. Kowey and G. V. Naccarelli, "Antiarrhythmic Drug Therapy: Where Do We Go From Here?," *Circulation* 149, no. 11 (2024): 801–803.

14. T. J. Moore, *Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster* (Simon & Schuster, 1995).

15. T. R. Fleming and D. L. DeMets, "Surrogate End Points in Clinical Trials: Are We Being Misled?," *Annals of Internal Medicine* 125, no. 7 (1996): 605–613.

16. J. N. Ruskin, "The Cardiac Arrhythmia Suppression Trial (CAST)," *New England Journal of Medicine* 321, no. 6 (1989): 386–388.

17. C. A. S. T. C. Investigators, "Preliminary Report: Effect of Encainide and Flecainide on Mortality in a Randomized Trial of Arrhythmia Suppression After Myocardial Infarction," *New England Journal of Medicine* 321, no. 6 (1989): 406–412.

18. C. M. Pratt and L. A. Moye, "The Cardiac Arrhythmia Suppression Trial: Background, Interim Results and Implications," *American Journal of Cardiology* 65, no. 4 (1990): 20–29.

19. R. L. Prentice, "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria," *Statistics in Medicine* 8, no. 4 (1989): 431–440.

20. L. Parast, P. Gilbert, and L. Wu, "Statistical Challenges in the Identification, Validation, and Use of Surrogate Markers," 2022, https://www.birs.ca/cmo-workshops/2022/22w5184/report22w5184.pdf.

21. CRAN, "The Comprehensive R Archive Network," 2024, https://cran.r-project.org.

22. M. R. Elliott, "Surrogate Endpoints in Clinical Trials," *Annual Review of Statistics and Its Application* 10 (2023): 75–96.

23. L. Parast, L. Tian, T. Cai, and L. Palaniappan, "Statistical Methods to Evaluate Surrogate Markers," *Medical Care* 62, no. 2 (2024): 102–108.

24. P. B. Gilbert, Y. Fong, N. S. Hejazi, et al., "Four Statistical Frameworks for Assessing an Immune Correlate of Protection (Surrogate Endpoint) From a Randomized, Controlled, Vaccine Efficacy Trial," *Vaccine* 42, no. 9 (2024): 2181–2190.

25. A. Alonso, V. Elst, G. Molenberghs, M. Buyse, and T. Burzykowski, "An Information-Theoretic Approach for the Evaluation of Surrogate Endpoints Based on Causal Inference," *Biometrics* 72, no. 3 (2016): 669–677.

26. A. Alonso and G. Molenberghs, "Surrogate Marker Evaluation From an Information Theory Perspective," *Biometrics* 63, no. 1 (2007): 180–186.

27. A. Alonso, G. Molenberghs, T. Burzykowski, et al., "Prentice's Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints," *Biometrics* 60, no. 3 (2004): 724–728.

28. A. Conlon, J. Taylor, Y. Li, K. Diaz-Ordaz, and M. Elliott, "Links Between Causal Effects and Causal Association for Surrogacy Evaluation in a Gaussian Setting," *Statistics in Medicine* 36, no. 27 (2017): 4243–4265.

29. F. Stijven, A. Alonso, and G. Molenberghs, "Proportion of Treatment Effect Explained: An Overview of Interpretations," *Statistical Methods in Medical Research* 33, no. 7 (2024): 1278–1296.

30. L. S. Freedman, B. I. Graubard, and A. Schatzkin, "Statistical Validation of Intermediate Endpoints for Chronic Diseases," *Statistics in Medicine* 11, no. 2 (1992): 167–178.

31. D. Lin, T. Fleming, and V. De Gruttola, "Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker," *Statistics in Medicine* 16, no. 13 (1997): 1515–1527.

32. Y. Wang and J. M. Taylor, "A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker," *Biometrics* 58, no. 4 (2002): 803–812.

33. J. M. Robins and S. Greenland, "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology* 3, no. 2 (1992): 143–155.

34. T. J. VanderWeele, "Surrogate Measures and Consistent Surrogates," *Biometrics* 69, no. 3 (2013): 561–565.

35. S. Athey, R. Chetty, G. W. Imbens, and H. Kang, "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *National Bureau of Economic Research* (2019): w26463.

36. M. M. Joffe and T. Greene, "Related Causal Frameworks for Surrogate Outcomes," *Biometrics* 65, no. 2 (2009): 530–538.

37. J. M. Taylor, Y. Wang, and R. Thiébaut, "Counterfactual Links to the Proportion of Treatment Effect Explained by a Surrogate Marker," *Biometrics* 61, no. 4 (2005): 1102–1111.

38. D. Ghosh, "Semiparametric Inference for Surrogate Endpoints With Bivariate Censored Data," *Biometrics* 64, no. 1 (2008): 149–156.

39. C. E. Frangakis and D. B. Rubin, "Principal Stratification in Causal Inference," *Biometrics* 58, no. 1 (2002): 21–29.

40. P. B. Gilbert and M. G. Hudgens, "Evaluating Candidate Principal Surrogate Endpoints," *Biometrics* 64, no. 4 (2008): 1146–1154.

41. E. E. Gabriel and D. Follmann, "Augmented Trial Designs for Evaluation of Principal Surrogates," *Biostatistics* 17, no. 3 (2016): 453–467.

42. Y. Huang and P. B. Gilbert, "Comparing Biomarkers as Principal Surrogate Endpoints," *Biometrics* 67, no. 4 (2011): 1442–1451.

43. A. S. Conlon, J. M. Taylor, and M. R. Elliott, "Surrogacy Assessment Using Principal Stratification When Surrogate and Outcome Measures Are Multivariate Normal," *Biostatistics* 15, no. 2 (2014): 266–283.

44. E. K. Roberts, M. R. Elliott, and J. M. Taylor, "Solutions for Surrogacy Validation With Longitudinal Outcomes for a Gene Therapy," *Biometrics* 79, no. 3 (2023): 1840–1852.

45. Y. Li, J. M. Taylor, and M. R. Elliott, "A Bayesian Approach to Surrogacy Assessment Using Principal Stratification in Clinical Trials," *Biometrics* 66, no. 2 (2010): 523–531.

46. E. K. Roberts, M. R. Elliott, and J. M. Taylor, "Surrogacy Validation for Time-To-Event Outcomes With Illness-Death Frailty Models," *Biometrical Journal* 66, no. 1 (2024): 2200324.

47. A. Conlon, J. Taylor, and M. Elliott, "Surrogacy Assessment Using Principal Stratification and a Gaussian Copula Model," *Statistical Methods in Medical Research* 26, no. 1 (2017): 88–107.

48. J. M. Taylor, A. S. Conlon, and M. R. Elliott, "Surrogacy Assessment Using Principal Stratification With Multivariate Normal and Gaussian Copula Models," *Clinical Trials* 12, no. 4 (2015): 317–322.

49. M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys, "The Validation of Surrogate Endpoints in Meta-Analyses of Randomized Experiments," *Biostatistics* 1, no. 1 (2000): 49–67.

50. M. J. Daniels and M. D. Hughes, "Meta-Analysis for the Evaluation of Potential Surrogate Markers," *Statistics in Medicine* 16, no. 17 (1997): 1965–1982.

51. G. Molenberghs, T. Burzykowski, A. Alonso, P. Assam, A. Tilahun, and M. Buyse, "The Meta-Analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials," *Journal of Statistical Planning and Inference* 138, no. 2 (2008): 432–449.

52. D. Ghosh, J. M. Taylor, and D. J. Sargent, "Meta-Analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling," *Biometrics* 68, no. 1 (2012): 226–233.

53. M. H. Gail, R. Pfeiffer, H. C. Van Houwelingen, and R. J. Carroll, "On Meta-Analytic Assessment of Surrogate Outcomes," *Biostatistics* 1, no. 3 (2000): 231–246.

54. E. E. Gabriel, M. C. Sachs, and M. E. Halloran, "Evaluation and Comparison of Predictive Individual-Level General Surrogates," *Biostatistics* 19, no. 3 (2018): 307–324.

55. A. Alonso and G. Molenberghs, "Evaluating Time to Cancer Recurrence as a Surrogate Marker for Survival From an Information Theory Perspective," *Statistical Methods in Medical Research* 17, no. 5 (2008): 497–504.

56. Y. Li, J. M. Taylor, M. R. Elliott, and D. J. Sargent, "Causal Assessment of Surrogacy in a Meta-Analysis of Colorectal Cancer Trials," *Biostatistics* 12, no. 3 (2011): 478–492.

57. L. Parast, M. M. McDermott, and L. Tian, "Robust Estimation of the Proportion of Treatment Effect Explained by Surrogate Marker Information," *Statistics in Medicine* 35, no. 10 (2016): 1637–1653.

58. M. Yp and B. W. Silverman, "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 61 (1982): 405–415.

59. A. Ullah and A. Pagan, *Nonparametric Econometrics* (Cambridge University Press, 1999).

60. R. R. Zhou, S. D. Zhao, and L. Parast, "Estimation of the Proportion of Treatment Effect Explained by a High-Dimensional Surrogate," *Statistics in Medicine* 41, no. 12 (2022): 2227–2246.

61. D. Agniel, B. P. Hejblum, R. Thiébaut, and L. Parast, "Doubly Robust Evaluation of High-Dimensional Surrogate Markers," *Biostatistics* 24, no. 4 (2023): 985–999.

62. V. d. M. J. Laan, E. C. Polley, and A. E. Hubbard, "Super Learner," *Statistical Applications in Genetics and Molecular Biology* 6, no. 1 (2007).

63. N. Meinshausen, "Relaxed Lasso," *Computational Statistics & Data Analysis* 52, no. 1 (2007): 374–393.

64. D. Agniel and L. Parast, "Evaluation of Longitudinal Surrogate Markers," *Biometrics* 77, no. 2 (2021): 477–489.

65. J. Xu and S. L. Zeger, "The Evaluation of Multiple Surrogate Endpoints," *Biometrics* 57, no. 1 (2001): 81–87.

66. U. G. Dafni and A. A. Tsiatis, "Evaluating Surrogate Markers of Clinical Outcome When Measured With Error," *Biometrics* 54, no. 4 (1998): 1445–1462.

67. S. Sarkar and Y. Qu, "Quantifying the Treatment Effect Explained by Markers in the Presence of Measurement Error," *Statistics in Medicine* 26, no. 9 (2007): 1955–1963.

68. W. Li and Y. Qu, "Adjustment for the Measurement Error in Evaluating Biomarkers," *Statistics in Medicine* 29, no. 22 (2010): 2338–2346.

69. L. Parast, T. P. Garcia, R. L. Prentice, and R. J. Carroll, "Robust Methods to Correct for Measurement Error When Evaluating a Surrogate Marker," *Biometrics* 78, no. 1 (2022): 9–23.

70. J. R. Cook and L. A. Stefanski, "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association* 89, no. 428 (1994): 1314–1328.

71. R. J. Carroll, H. Küchenhoff, F. Lombard, and L. A. Stefanski, "Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models," *Journal of the American Statistical Association* 91, no. 433 (1996): 242–250.

72. L. Parast, T. Cai, and L. Tian, "Evaluating Surrogate Marker Information Using Censored Data," *Statistics in Medicine* 36, no. 11 (2017): 1767–1782.

73. D. M. Dabrowska, "Non-Parametric Regression With Censored Survival Time Data," *Scandinavian Journal of Statistics* 13, no. 3 (1987): 181–197.

74. R. Beran, *Nonparametric Regression With Randomly Censored Survival Data*Technical report (University of California, 1981).

75. L. Parast, T. Cai, and L. Tian, "Evaluating Multiple Surrogate Markers With Censored Data," *Biometrics* 77, no. 4 (2021): 1315–1327.

76. L. Parast, L. Tian, and T. Cai, "Assessing the Value of a Censored Surrogate Outcome," *Lifetime Data Analysis* 26 (2020): 245–265.

77. D. Agniel and L. Parast, "Robust Evaluation of Longitudinal Surrogate Markers With Censored Data," 2024, *arXiv Preprint arXiv:2402.16969*.

78. D. Ghosh, "On Assessing Surrogacy in a Single Trial Setting Using a Semicompeting Risks Paradigm," *Biometrics* 65, no. 2 (2009): 521–529.

79. E. E. Gabriel and P. B. Gilbert, "Evaluating Principal Surrogate Endpoints With Time-To-Event Data Accounting for Time-Varying Treatment Efficacy," *Biostatistics* 15, no. 2 (2014): 251–265.

80. E. E. Gabriel, M. C. Sachs, and P. B. Gilbert, "Comparing and Combining Biomarkers as Principle Surrogates for Time-To-Event Clinical Endpoints," *Statistics in Medicine* 34, no. 3 (2015): 381–395.

81. L. Parast, T. Cai, and L. Tian, "Using a Surrogate With Heterogeneous Utility to Test for a Treatment Effect," *Statistics in Medicine* 42, no. 1 (2023): 68–88.

82. L. Parast, T. Cai, and L. Tian, "Testing for Heterogeneity in the Utility of a Surrogate Marker," *Biometrics* 79, no. 2 (2023): 799–810.

83. E. K. Roberts, M. R. Elliott, and J. M. Taylor, "Incorporating Baseline Covariates to Validate Surrogate Endpoints With a Constant Biomarker Under Control Arm," *Statistics in Medicine* 40, no. 29 (2021): 6605–6618.

84. M. R. Elliott, A. S. Conlon, Y. Li, N. Kaciroti, and J. M. Taylor, "Surrogacy Marker Paradox Measures in Meta-Analytic Settings," *Biostatistics* 16, no. 2 (2015): 400–412.

85. M. C. Sachs, E. E. Gabriel, A. Crippa, and M. J. Daniels, "Flexible Evaluation of Surrogacy in Platform Studies," *Biostatistics* 25, no. 1 (2024): 220–236.

86. L. Parast, L. Tian, and T. Cai, "Assessing Heterogeneity in Surrogacy Using Censored Data," *Statistics in Medicine* 43, no. 17 (2024): 3184–3209.

87. R. Knowlton, L. Tian, and L. Parast, "A General Framework to Assess Complex Heterogeneity in the Utility of a Surrogate Marker," *Statistics in Medicine* 44, no. 5 (2025): e70001.

88. B. L. Price, P. B. Gilbert, and V. D. M. J. Laan, "Estimation of the Optimal Surrogate Based on a Randomized Trial," *Biometrics* 74, no. 4 (2018): 1271–1281.

89. X. Wang, L. Parast, L. Tian, and T. Cai, "Model-Free Approach to Quantifying the Proportion of Treatment Effect Explained by a Surrogate Marker," *Biometrika* 107, no. 1 (2020): 107–122.

90. X. Wang, T. Cai, L. Tian, F. Bourgeois, and L. Parast, "Quantifying the Feasibility of Shortening Clinical Trial Duration Using Surrogate Markers," *Statistics in Medicine* 40, no. 28 (2021): 6321–6343.

91. X. Wang, L. Parast, L. Han, L. Tian, and T. Cai, "Robust Approach to Combining Multiple Markers to Improve Surrogacy," *Biometrics* 79, no. 2 (2023): 788–798.

92. H. Chen, Z. Geng, and J. Jia, "Criteria for Surrogate End Points," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 69, no. 5 (2007): 919–932.

93. Y. Yin, L. Liu, Z. Geng, and P. Luo, "Novel Criteria to Exclude the Surrogate Paradox and Their Optimalities," *Scandinavian Journal of Statistics* 47, no. 1 (2020): 84–103.

94. E. Hsiao, L. Tian, and L. Parast, "Avoiding the Surrogate Paradox: An Empirical Framework for Assessing Assumptions," *Under Review* (2024).

95. J. Zhou, X. Jiang, H. A. Xia, B. P. Hobbs, and P. Wei, "Landmark Mediation Survival Analysis Using Longitudinal Surrogate," *Frontiers in Oncology* 12 (2023): 999324.

96. R. L. Prentice, "Surrogate and Mediating Endpoints: Current Status and Future Directions," *JNCI Journal of the National Cancer Institute* 101, no. 4 (2009): 216–217.

97. E. A. Croce, L. Parast, D. Bhavnani, and E. C. Matsui, "Lower Socioeconomic Status May Help Explain Racial Disparities in Asthma and Atopic Dermatitis Prevalence: A Mediation Analysis," *Journal of Allergy and Clinical Immunology* 153, no. 4 (2024): 1140–1147.

98. E. Agyemang, A. S. Magaret, S. Selke, C. Johnston, L. Corey, and A. Wald, "Herpes Simplex Virus Shedding Rate: Surrogate Outcome for Genital Herpes Recurrence Frequency and Lesion Rates, and Phase 2 Clinical Trials End Point for Evaluating Efficacy of Antivirals," *Journal of Infectious Diseases* 218, no. 11 (2018): 1691–1699.

99. T. Sprenger, L. Kappos, E. W. Radue, et al., "Association of Brain Volume Loss and Long-Term Disability Outcomes in Patients With Multiple Sclerosis Treated With Teriflunomide," *Multiple Sclerosis Journal* 26, no. 10 (2020): 1207–1216.

100. L. M. Ruilope, R. Agarwal, S. D. Anker, et al., "Blood Pressure and Cardiorenal Outcomes With Finerenone in Chronic Kidney Disease in Type 2 Diabetes," *Hypertension* 79, no. 12 (2022): 2685–2695.

101. C. Dromain, M. Pavel, M. Ronot, et al., "Response Heterogeneity as a New Biomarker of Treatment Response in Patients With Neuroendocrine Tumors," *Future Oncology* 19, no. 32 (2023): 2171–2183.

102. B. S. Blette, J. Moutchia, N. Al-Naamani, et al., "Is Low-Risk Status a Surrogate Outcome in Pulmonary Arterial Hypertension? An Analysis of Three Randomised Trials," *Lancet Respiratory Medicine* 11, no. 10 (2023): 873–882.

103. M. R. Elliott, Y. Li, and J. M. Taylor, "Accommodating Missingness When Assessing Surrogacy via Principal Stratification," *Clinical Trials* 10, no. 3 (2013): 363–377.

104. Z. Zhang and L. Wang, "Methods for Mediation Analysis With Missing Data," *Psychometrika* 78 (2013): 154–184.

105. W. Li and X. H. Zhou, "Identifiability and Estimation of Causal Mediation Effects With Missing Data," *Statistics in Medicine* 36, no. 25 (2017): 3948–3965.

106. X. Wang, L. Parast, L. Tian, and T. Cai, "Towards Optimal Use of Surrogate Markers to Improve Power," 2022, *arXiv Preprint arXiv:2209.08414*.

107. L. Parast and J. Bartroff, "Group Sequential Testing of a Treatment Effect Using a Surrogate Marker," *Biometrics* 80, no. 4 (2024): ujae108.

108. L. Parast, T. Cai, and L. Tian, "Using a Surrogate Marker for Early Testing of a Treatment Effect," *Biometrics* 75, no. 4 (2019): 1253–1263.

109. E. Bareinboim and J. Pearl, "A General Algorithm for Deciding Transportability of Experimental Results," *Journal of Causal Inference* 1, no. 1 (2013): 107–134.

110. M. Daniels, C. Frangakis, V. Charu, and D. Ghosh, "University of Pennsylvania 7th Annual Conference on Statistical Issues in Clinical Trials: Current Issues Regarding the Use of Biomarkers and Surrogate Endpoints in Clinical Trials (Morning Panel Discussion)," *Clinical Trials* 12, no. 4 (2015): 323–332.

111. W. M. Kouw and M. Loog, "A Review of Domain Adaptation Without Target Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 3 (2019): 766–785.

112. W. M. Kouw and M. Loog, "An Introduction to Domain Adaptation and Transfer Learning," 2018, *arXiv Preprint arXiv:1812.11806*.

113. I. Degtiar and S. Rose, "A Review of Generalizability and Transportability," *Annual Review of Statistics and Its Application* 10, no. 1 (2023): 501–524.

114. J. D. Wallach, S. Yoon, H. Doernberg, et al., "Associations Between Surrogate Markers and Clinical Outcomes for Nononcologic Chronic Disease Treatments," *JAMA* 331, no. 19 (2024): 1646–1654.