Biostatistics (2023) **24**, 4, *pp.* 985–999
https://doi.org/10.1093/biostatistics/kxac020
Advance Access publication on July 6, 2022

Doubly robust evaluation of high-dimensional surrogate markers

DENIS AGNIEL*

RAND Corporation, 1776 Main St. Santa Monica, CA, 90401, USA
dagniel@rand.org

BORIS P. HEJBLUM

Univ. Bordeaux, INSERM, INRIA, BPH, U1219, SISTM, F-33000 Bordeaux, France and Vaccine Research Institute, F-94000 Créteil, France

RODOLPHE THIÉBAUT

Univ. Bordeaux, INSERM, INRIA, BPH, U1219, SISTM, F-33000 Bordeaux, France, CHU de Bordeaux, Service d'Information médicale, F-33000 Bordeaux, France and Vaccine Research Institute, F-94000 Créteil, France

LAYLA PARAST

University of Texas at Austin, Department of Statistics and Data Sciences, 3925 West Braker Lane, Austin, TX 78759, USA

SUMMARY

When evaluating the effectiveness of a treatment, policy, or intervention, the desired measure of efficacy may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily, quickly, or cheaply capture the effect of interest. Theory and methods for evaluating the strength of surrogate markers have been well studied in the context of a single surrogate marker measured in the course of a randomized clinical study. However, methods are lacking for quantifying the utility of surrogate markers when the dimension of the surrogate grows. We propose a robust and efficient method for evaluating a set of surrogate markers that may be high-dimensional. Our method does not require treatment to be randomized and may be used in observational studies. Our approach draws on a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference—namely, methods for robust estimation of the average treatment effect. This connection facilitates the use of modern methods for estimating treatment effects, using machine learning to estimate nuisance functions and relaxing the dependence on model specification. We demonstrate that our proposed approach performs well, demonstrate connections between our approach and certain mediation effects, and illustrate it by evaluating whether gene expression can be used as a surrogate for immune activation in an Ebola study.

Keywords: Average treatment effect estimation; High-dimensional data; Surrogate marker evaluation.

*To whom correspondence should be addressed.

1. INTRODUCTION

When evaluating the effectiveness of a treatment, policy, or intervention, the desired measure of effectiveness may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily, quickly, or cheaply capture the effect of interest. For example, when evaluating an intervention designed to delay dementia onset, the time required to observe enough dementia diagnoses is often very long, and surrogates that have been considered for intervention evaluation include mild cognitive impairment, adiponectin levels, neuroimages, amyloid plaques, and neurofibrillary tangles (Small, 2006; Teixeira *and others*, 2013). Similarly, in studies evaluating treatments to prevent diabetes, surrogate measures for diabetes onset have included changes in body weight, fasting plasma glucose, and hemoglobin A1c (Caveney and Cohen, 2011; Choi *and others*, 2011). And it is presently obvious that a surrogate marker for SARS-CoV-2 vaccine efficacy is required to develop the second generation of vaccines adapted to novel variants of concern (Karim, 2021).

Theory and methods for evaluating the strength of surrogate markers have been well studied in the context of a single surrogate marker measured in the course of one or more randomized clinical studies—see the detailed review in Joffe and Greene (2009). In particular, model-based approaches have been proposed for continuous (Alonso *and others*, 2004) and binary surrogates (Alonso *and others*, 2016) in meta-analysis and in the principal stratification framework (Gilbert and Hudgens, 2008). Robust nonparametric methods have been proposed in Parast *and others* (2020) and Parast *and others* (2016). However, fully nonparametric methods are not available or not reliable when the number of markers is more than one or two. In these cases, a parametric approach to evaluate a high-dimensional surrogate may be considered as proposed in Zhou *and others* (2022), which requires the correct specification of two high-dimensional linear models. Alternatively, an initial model may be used to reduce the dimensionality of the surrogate (Agniel and Parast, 2021; Parast *and others*, 2016). However, the mis-specification of the parametric models or an inappropriate initial model may produce badly biased estimates of the utility of the surrogate.

In this work, we propose a robust and efficient method for evaluating multiple surrogate markers, with particular attention to the possibility that surrogates may be high-dimensional. Our approach revives the spirit of the estimator in Freedman *and others* (1992), itself based on the pioneering work of Prentice (1989) to draw on a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference: namely, methods for estimating the average treatment effect. This connection does not appear to have been made before in the surrogate marker literature, despite the extensive connections made between mediation and surrogacy (Taylor *and others*, 2005; Joffe and Greene, 2009). Making this connection facilitates the use of all of the machinery now available to robustly and efficiently estimate average treatment effects. Specifically, we take advantage of state-of-the-art methods for incorporating flexible machine learning and/or sparse high-dimensional models into the estimation of treatment effects. These methods based estimation on the efficient influence function of key quantities and use sample splitting to ensure convenient asymptotic distributions and inference while putting minimal restrictions on what types of estimators may be used. Furthermore, we show that our proposed approach has connections to certain mediation estimands when the assumptions underlying mediation analysis are satisfied.

One key benefit of our approach is that it does not require randomization of treatment, which has been required by previous nonparametric approaches to evaluating surrogate markers (Agniel and Parast, 2021; Parast *and others*, 2016). Assessments of surrogate markers are not solely limited to studies of randomized treatment (Obirikorang *and others*, 2012). In fact, using surrogate outcomes based on complex observational data is finding purchase in all corners of science (Wang *and others*, 2020). In Section 6, we give an example where researchers are interested in studying whether gene expression can be used as a surrogate for immune activation in an observational setting. Our approach exploits insights from observational causal inference to estimate all relevant quantities and thus is equally applicable in both observational and randomized settings.

The structure of the rest of the article is as follows. In Section 2, we lay out the notation and the setting in which we are working, and we motivate the importance of evaluating surrogate markers in light of recent advances in the surrogate marker literature. In Section 3, we detail assumptions necessary for identifying and interpreting parameters of interest. We discuss the identification, estimation, and the asymptotic behavior of our estimator in Section 4 and discuss variance estimation and inference. We evaluate the performance of the proposed approach using a simulation study in Section 5, and in Section 6, we investigate whether gene expression could be used as a surrogate for immune response to Ebola infection. We give final remarks and draw connections to other methods for evaluating surrogate markers in Section 7.

2. Evaluation of surrogate markers

2.1. *Notation*

Let A denote a binary treatment, and let the primary outcome of the study be Y. Let there be a vector $\mathbf{S}$ of potential surrogate information, and let $\mathbf{X}$ be a vector of pretreatment covariates. The primary quantity of interest is the treatment effect on the outcome $\Delta = \mathbb{E}\{Y^{(1)} - Y^{(0)}\}$, where $Y^{(a)}$ is the potential or counterfactual outcome that would have been observed if treatment were $A = a$, possibly contrary to fact. Similarly, let $\mathbf{S}^{(a)}$ be the potential/counterfactual value the vector of surrogates would take if $A = a$. Let the data observed in the current study be $\mathcal{O} = (\mathbf{X}_i, A_i, \mathbf{S}_i, Y_i)_{i=1,\ldots,n}$, n iid realizations of $\mathbf{O} = (\mathbf{X}, A, \mathbf{S}, Y)$.

2.2. *Importance of quantifying surrogate strength*

Understanding the strength of a potential surrogate is important for many reasons. In practice, information about surrogate strength will inform decisions regarding whether to measure the surrogate in a future study (especially if it is costly or invasive to measure) and/or whether to use the surrogate to assess treatment effectiveness in a future study. In addition, a number of novel statistical methods for using surrogates in future studies can only be applied when the surrogate is strong. For example, Parast *and others* (2019) propose a robust nonparametric procedure to test for a treatment effect using surrogate marker information measured prior to the end of the study in a time-to-event outcome setting, but they rely on the assumption that the surrogate is sufficiently strong. As another example, Price *and others* (2018) propose constructing a function $\psi_a(\cdot)$ of the surrogates that leads to some optimality properties, but it can be shown that this method should only be used with a strong surrogate.

Specifically, they construct an optimal transformation of the surrogates $\psi_a(\cdot)$ in terms of minimizing the following mean squared error

$$\mathcal{M}_\psi = \mathbb{E}\left[e_1(\mathbf{X})\left\{Y^{(1)} - \psi_1(\mathbf{X}, \mathbf{S}^{(1)})\right\}^2\right] + \mathbb{E}\left[e_0(\mathbf{X})\left\{Y^{(0)} - \psi_0(\mathbf{X}, \mathbf{S}^{(0)})\right]^2\right], \tag{2.1}$$

with $e_a(\mathbf{x}) = \mathbb{P}(A = a|\mathbf{X})$, under the constraint that it satisfies the so-called Prentice definition:

$$\mathbb{E}(Y^{(1)} - Y^{(0)}) = 0 \text{ if and only if } \mathbb{E}\left\{\psi_1(\mathbf{X}, \mathbf{S}^{(1)}) - \psi_0(\mathbf{X}, \mathbf{S}^{(0)})\right\} = 0.$$

The optimal transformations in this case are shown to be $\psi_a(\mathbf{x}, \mathbf{s}) = \mathbb{E}\{Y^{(a)}|\mathbf{X} = \mathbf{x}, \mathbf{S}^{(a)} = \mathbf{s}\}$. In addition to this appealing optimality property, this proposal also resolves the so-called "surrogate paradox." The paradox states that the treatment could have a positive causal effect on a univariate surrogate S which could have a positive correlation with the outcome, but the treatment could yet have a negative effect on the outcome. The Price surrogate resolves the surrogate paradox by definition. The treatment effect on the transformed surrogate is by definition equivalent to the treatment effect on the outcome: $\mathbb{E}\{\psi_1(\mathbf{X}, \mathbf{S}^{(1)})\} - \mathbb{E}\{\psi_0(\mathbf{X}, \mathbf{S}^{(0)})\} = \Delta$.

However, the Price surrogate's resolution of the surrogate paradox is in some sense too good. The surrogate paradox is thus resolved for every potential set of surrogates, good and bad. Even if the surrogates $\mathbf{S}$ are completely unrelated to the outcome Y, $\mathbb{E}\{\psi_1(\mathbf{X}, \mathbf{S}^{(1)})\} - \mathbb{E}\{\psi_0(\mathbf{X}, \mathbf{S}^{(0)})\} = \Delta$. In fact, the power to detect a treatment effect in a future study using $\psi_a(\cdot)$ actually increases as the surrogate becomes weaker (explaining less of the treatment effect), if the Prentice definition does not hold. Consider the toy example depicted in Figure 1 (see Appendix A of the Supplementary material available at *Biostatistics* online for details of the data generation). As the strength of the surrogate increases, $\psi_a(\cdot)$ becomes more like the true outcome $Y_i^{(a)}$. As the surrogate becomes weaker, $\psi_a(\cdot)$ becomes more like $\bar{Y}_a$ the mean outcome in treatment group $A = a$ from the first study. This means that a weak surrogate ensures that the treatment effect in the second study will be identical to the treatment effect in the first study because the distribution of $\psi_a(\cdot)$ will cluster closely around the estimate of $\mathbb{E}(Y^{(a)})$ from the first study (see specifically, the fourth panel of Figure 1). In this scenario, the second study is providing no new information about the treatment. Thus, the Price surrogate should only be used with a strong $\mathbf{S}$, and its use with a weak or possibly even moderately strong set of surrogates may not be advisable.

Therefore, for both decision-making purposes and to use statistical methods that take advantage of strong surrogates, methods are needed to rigorously quantify the strength of the proposed set of surrogates before using (functions of) the surrogates in practice. In the following section, we propose an efficient method for estimating the strength of a possibly high-dimensional set of surrogates in observational studies. Our aim is for this proposed method to complement existing work, like that of Parast *and others* (2019) and Price *and others* (2018), that rely on the availability of a strong surrogate but currently lack tools to assess surrogate strength, particularly in a high-dimensional surrogate setting.

2.3. *Quantity to evaluate surrogate strength*

In this section, we present a general method for evaluating the usefulness of a set of surrogates. This can be used to evaluate the suitability of any set of surrogates or surrogate transformations and is appropriate for $\mathbf{S}$ of any dimension. To evaluate the surrogates' usefulness, we use the proportion of treatment effect explained by $\mathbf{S}$ (PTE) defined as $R_{\mathbf{S}} = (\Delta - \Delta_{\mathbf{S}})/\Delta = 1 - \Delta_{\mathbf{S}}/\Delta$, with $\Delta_{\mathbf{S}}$ the residual treatment effect, or the treatment effect that remains after controlling for the surrogate information. This measure has been used frequently in the evaluation of surrogate markers, as far back as the pioneering works of Prentice (1989) and Freedman *and others* (1992) and more recently in Parast *and others* (2020, 2016) and Agniel and Parast (2021). If $\mathbf{S}^{(0)}, \mathbf{S}^{(1)}$ are independent of $Y^{(0)}, Y^{(1)}$ conditional on $\mathbf{X}$, then $\Delta_{\mathbf{S}} = \Delta$ and $R_{\mathbf{S}} = 0$. In contrast, if all of the treatment effects can be attributed to $\mathbf{S}$, then $\Delta_{\mathbf{S}} = 0$ and $R_{\mathbf{S}} = 1$.

In the more recent works using this PTE, $\Delta_{\mathbf{S}}$ is defined in terms of a particular reference distribution, e.g., taking $\Delta_{\mathbf{S}} = \mathbb{E}\{\psi_1(\mathbf{S}) - \psi_0(\mathbf{S})|A = 1\}$ to be the residual treatment effect among the treated group. This choice of reference distribution is often arbitrary and was not required by the older (model-based) approaches. To obviate the need for this choice of reference distribution and to take advantage of the extensive development in average treatment effect estimation (and the associated intuition and machinery that have been built up around it), we define $\Delta_{\mathbf{S}} = \mathbb{E}\{\psi_1(\mathbf{X}, \mathbf{S}) - \psi_0(\mathbf{X}, \mathbf{S})\}$, the average treatment effect conditional on the distribution of $\mathbf{S}^{(0)}$ and $\mathbf{S}^{(1)}$ both being equal to the distribution of $\mathbf{S}$ (all conditional on $\mathbf{X}$). Other choices of reference distribution could be used when there are substantive reasons to prefer them. When there are no interactions with treatment – $\mathbb{P}\{\psi_1(\mathbf{X}, \mathbf{S}) = \psi_0(\mathbf{X}, \mathbf{S})\} = 1$ – choice of reference distribution does not change $\Delta_{\mathbf{S}}$. See Section 7 for further connections in this vein to Agniel and Parast (2021) and similar recent methods.

This formulation of PTE also has deep connections to quantities important to mediation analysis without requiring some of the restrictive assumptions common in the mediation literature. In particular, we show in Appendix B of the Supplementary material available at *Biostatistics* online that $\Delta_{\mathbf{S}}$ is a function of conditional natural direct effects and that $\Delta - \Delta_{\mathbf{S}}$ is a function of conditional natural indirect effects

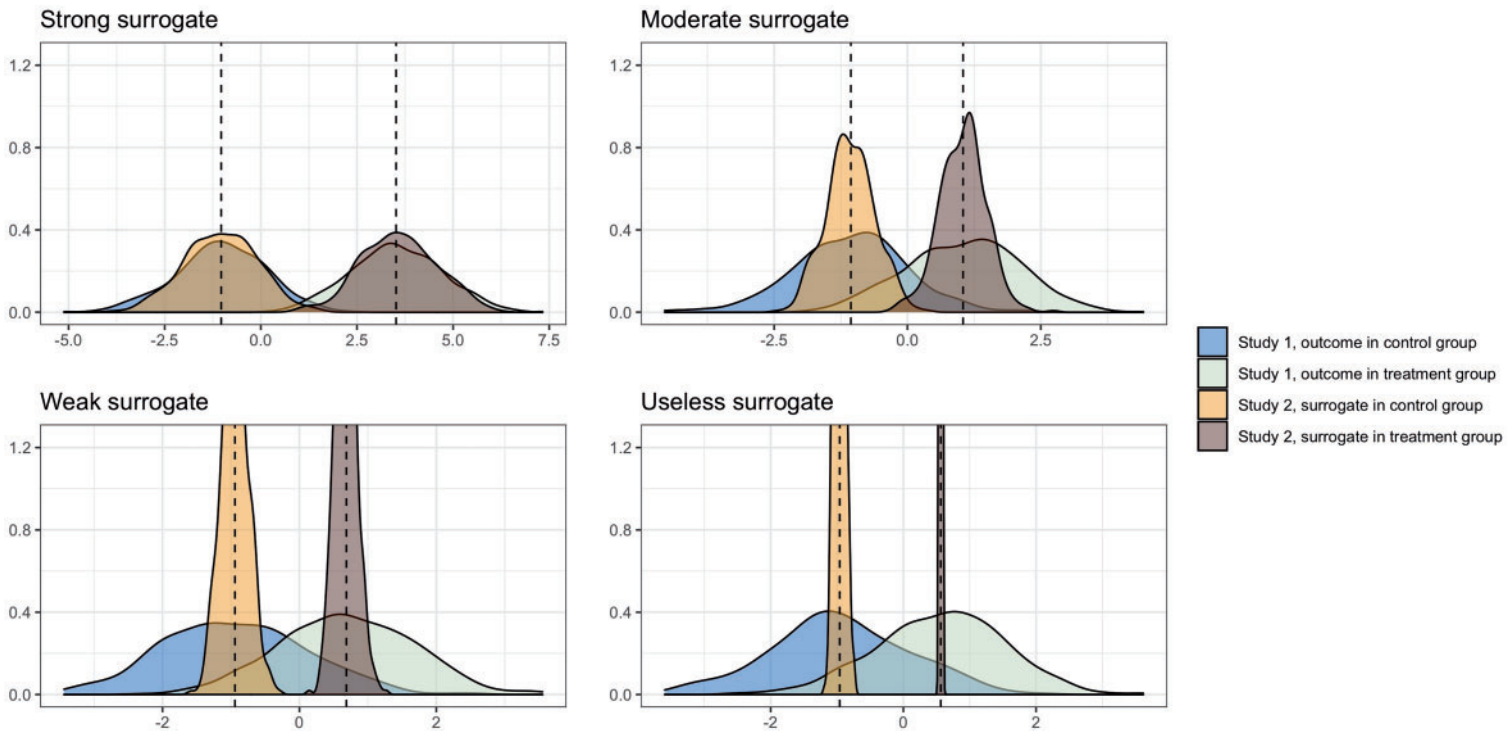

Fig. 1. Distribution of outcomes Y in an initial study (Study 1) and surrogate functions (approximating the outcome) ψ_1 (surrogate in the treatment group) and ψ_0 (surrogate in the control group) in a future study (Study 2) based on simulated data, arranged according to the strength of the surrogate marker—i.e., the capacity of the surrogate to explain the treatment effect. The wider distribution on the left corresponds to the control group in Study 1, the wider distribution on the right to the treated group in Study 1, the narrower distribution on the left to ψ_0 in Study 2, and the narrower distribution on the right to ψ_1 in Study 2. Vertical dotted lines correspond to the means of the treatment groups in the first study. See Appendix A of the Supplementary material available at *Biostatistics* online for simulation details. Surrogate functions ψ_0, ψ_1 were estimated via a correctly specified linear regression in the first study and then applied to the second study population. When the surrogate was strong, the distribution of the surrogate in the second study was quite close to the distribution of the outcome in the first study. However, when the surrogate was weak, the distribution of the surrogate in the second study clustered around the group mean from the first study.

(Joffe and Greene, 2009; VanderWeele, 2013), when the assumptions for identifying those effects are met. While many previous reviews of surrogate methods identify the causal effects framework for surrogate evaluation solely with mediation (Joffe and Greene, 2009; Conlon *and others*, 2017), we use causal tools without requiring all of the assumptions necessary for mediation. Importantly, we do not require that all confounders of the surrogate–outcome relationship are measured and included in the study because the aims of mediation and surrogate marker evaluation are different (VanderWeele, 2013). The aim of identifying a mediator is determining whether the effect of treatment operates through the mediator itself, e.g., through some biological pathway. Often, a good surrogate marker is similarly conceptualized as a variable through which the treatment operates, but this is not necessarily required; a variable can be a good surrogate if it captures the treatment effect on the outcome, even if the treatment effect does not operate through the variable itself (sometimes called a nonmechanistic correlate of protection; Plotkin and Gilbert, 2012).

3. ASSUMPTIONS

3.1. *Identifying assumptions*

We first require $\Delta \neq 0$, without which requirement the goals of identifying surrogate markers are practically and theoretically not meaningful. We further make the three typical assumptions of treatment effect estimation: consistency, positivity, and no unmeasured confounding. Specifically, we assume that the observed values of $\mathbf{S}$ and Y when $A = a$ are identical to the counterfactuals $\mathbf{S}^{(a)}$ and $Y^{(a)}$ such that $\mathbf{S} = \mathbf{S}^{(1)}A + \mathbf{S}^{(0)}(1-A)$ and $Y = Y^{(1)}A + Y^{(0)}(1-A)$. We furthermore assume that $\mathbf{X}$ contains all confounders of the effects of A on the surrogates and the outcome, such that the treatment A is as good as randomized conditional on the covariates $\mathbf{X}$: (A.1) $\{\mathbf{S}^{(0)}, \mathbf{S}^{(1)}, Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp A \,\big|\, \mathbf{X}$. In addition, we assume two forms of positivity, which ensures that individuals in the two study arms are not too different from one another. First, the usual positive probability of receiving either treatment for some $\epsilon_1 > 0$, $\mathbb{P}\{\epsilon_1 < e_1(\mathbf{X}) < 1 - \epsilon_1\} = 1$, and a related assumption that further conditions on the surrogates, $\mathbb{P}\{\epsilon_2 < \pi_1(\mathbf{X}, \mathbf{S}) < 1 - \epsilon_2\} = 1$, for some $\epsilon_2 > 0$, where $\pi_1(\mathbf{S}, \mathbf{X}) = \mathbb{P}(A = 1|\mathbf{S}, \mathbf{X})$. Notably, letting $f_{\mathbf{U}}(\mathbf{u})$ be the density of the random variable $\mathbf{U}$ evaluated at $\mathbf{u}$, because $\pi_1(\mathbf{x}, \mathbf{s}) = f_{\mathbf{S}}(\mathbf{s}|\mathbf{X} = \mathbf{x}, A = 1)e_1(\mathbf{x})/f_{\mathbf{S}}(\mathbf{s}|\mathbf{X} = \mathbf{x}) = f_{\mathbf{S}^{(1)}}(\mathbf{s}|\mathbf{X} = \mathbf{x})e_1(\mathbf{x})/\left[f_{\mathbf{S}^{(1)}}(\mathbf{s}|\mathbf{X} = \mathbf{x})e_1(\mathbf{x}) + f_{\mathbf{S}^{(0)}}(\mathbf{s}|\mathbf{X} = \mathbf{x})e_0(\mathbf{x})\right]$, these two positivity conditions ensure that the conditional distribution of the counterfactual surrogates under treatment and control cannot be too different from one another, i.e., ensuring overlap. When the treatment has a large effect on $\mathbf{S}$, this additional overlap requirement may be suspect—e.g., there may be some values of $\mathbf{s}$ such that $f_{\mathbf{S}^{(1)}}(\mathbf{s}|\mathbf{X} = \mathbf{x})$ approaches 0 and thus $\pi_1(\mathbf{x}, \mathbf{s})$ also approaches 0. We discuss a generalization of our approach that removes this overlap condition in Appendix C of the Supplementary material available at *Biostatistics* online.

3.2. *Required assumptions to ensure interpretation of $R_{\mathbf{S}}$ as a proportion*

The interpretation of $R_{\mathbf{S}}$ as the PTE depends on it actually being a proportion, lying between 0 and 1. We adapt conditions in Wang and Taylor (2002) and Agniel and Parast (2021) which ensure that $0 \leq R_{\mathbf{S}} \leq 1$. These conditions are as follows (assuming without loss of generality that $\Delta > 0$): (A.2) $\int \{e_1(\mathbf{x})\psi_1(\mathbf{x}, \mathbf{s}) + e_0(\mathbf{x})\psi_0(\mathbf{x}, \mathbf{s})\}\, \mathrm{d}\left\{F_{\mathbf{X},\mathbf{S}^{(1)}}(\mathbf{x}, \mathbf{s}) - F_{\mathbf{X},\mathbf{S}^{(0)}}(\mathbf{x}, \mathbf{s})\right\} \geq 0$; and (A.3) $\int \{\psi_1(\mathbf{x}, \mathbf{s}) - \psi_0(\mathbf{x}, \mathbf{s})\}\, \mathrm{d}F_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s}) \geq 0$. The condition (A.3) ensures that $\Delta_S \geq 0$, i.e., that the residual treatment effect is in the same direction as the overall treatment effect Δ. Condition (A.2), which ensures that $\Delta \geq \Delta_S$, requires that, roughly speaking, a propensity-weighted mixture of the two conditional mean functions $e_1(\mathbf{x})\psi_1(\mathbf{x}, \mathbf{s}) + e_0(\mathbf{x})\psi_0(\mathbf{x}, \mathbf{s})$ is larger when $\mathbf{s}$ takes values from the distribution of the counterfactual surrogates under treatment than if it took values from the distribution under control. These two conditions are analogs of conditions (A6) and (A7) in Agniel and Parast (2021). See Appendix D of the Supplementary material available at *Biostatistics* online for an alternative approach that may be

considered if these conditions are not or unlikely to be met, though the alternative approach lacks the interpretability and connections to previous literature of the proposed approach.

4. Identification, estimation, and inference

4.1. *Identifiability*

In this section, we show that the effects of interest are identifiable, propose a robust and efficient estimation procedure, and describe the asymptotic properties of our proposed estimators. The effects of interest may be identified as average treatment effects identifiable from the data, one of which conditions on the surrogates ($\Delta_{\mathbf{S}}$) and one which does not (Δ). In particular, the residual treatment effect may be identified as

$$\Delta_{\mathbf{S}} = \mathbb{E}\left\{\mu_1(\mathbf{X},\mathbf{S}) - \mu_0(\mathbf{X},\mathbf{S})\right\} = \mathbb{E}\left[\frac{AY}{\pi_1(\mathbf{X},\mathbf{S})} - \frac{(1-A)Y}{\pi_0(\mathbf{X},\mathbf{S})}\right] \tag{4.2}$$

$$= \mathbb{E}\left[\frac{AY - \{A-\pi_1(\mathbf{X},\mathbf{S})\}\mu_1(\mathbf{X},\mathbf{S})}{\pi_1(\mathbf{X},\mathbf{S})} - \frac{(1-A)Y - \{1-A-\pi_0(\mathbf{X},\mathbf{S})\}\mu_0(\mathbf{X},\mathbf{S})}{\pi_0(\mathbf{X},\mathbf{S})}\right], \tag{4.3}$$

where $\mu_a(\mathbf{x},\mathbf{s}) = \mathbb{E}(Y|\mathbf{X}=\mathbf{x}, A=a, \mathbf{S}=\mathbf{s})$. The result in the first equality of (4.2) follows from the definition of $\Delta_{\mathbf{S}}$ and the fact that $\psi_a(\mathbf{X},\mathbf{S}^{(a)}) = \mu_a(\mathbf{X},\mathbf{S})$ because of (A.1), and the other results follow shortly thereafter, following familiar paths as arguments for the identification of the average treatment effect in other contexts. These results show that the residual treatment effect may be identified without knowledge of the mean function for the outcome via the second equality in (4.2) and gives an augmented inverse probability weighting (Bang and Robins, 2005) version of the estimand (4.3). Δ may also be identified, following similar standard arguments and using similar functionals with $\mu_a(\mathbf{X},\mathbf{S})$ replaced by $m_a(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}=\mathbf{x}, A=a)$ and $\pi_a(\mathbf{X},\mathbf{S})$ replaced by $e_a(\mathbf{X})$.

4.2. *Proposed estimation*

This identification of Δ and $\Delta_{\mathbf{S}}$ in terms of average treatment effects allows us to take advantage of the rich literature on robust estimation of these quantities. We propose to estimate $\Delta_{\mathbf{S}}$ and Δ as

$$\widehat{\Delta}_{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\left[\frac{A_iY_i - \{A_i-\widehat{\pi}_1(\mathbf{X}_i,\mathbf{S}_i)\}\widehat{\mu}_1(\mathbf{X}_i,\mathbf{S}_i)}{\widehat{\pi}_1(\mathbf{X}_i,\mathbf{S}_i)} - \frac{(1-A_i)Y_i - \{1-A_i-\widehat{\pi}_0(\mathbf{X}_i,\mathbf{S}_i)\}\widehat{\mu}_0(\mathbf{X}_i,\mathbf{S}_i)}{\widehat{\pi}_0(\mathbf{X}_i,\mathbf{S}_i)}\right]$$

$$\widehat{\Delta} = n^{-1}\sum_{i=1}^{n}\left[\frac{A_iY_i - \{A_i-\widehat{e}_1(\mathbf{X}_i)\}\widehat{m}_1(\mathbf{X}_i)}{\widehat{e}_1(\mathbf{X}_i)} - \frac{(1-A_i)Y_i - \{1-A_i-\widehat{e}_0(\mathbf{X}_i)\}\widehat{m}_0(\mathbf{X}_i)}{\widehat{e}_0(\mathbf{X}_i)}\right]$$

and thus, estimate $R_{\mathbf{S}}$ as $\widehat{R}_{\mathbf{S}} = 1-\widehat{\Delta}_{\mathbf{S}}/\widehat{\Delta}$, where we leave estimation of the components: $\pi_1(\mathbf{X},\mathbf{S})$, $\mu_1(\mathbf{X},\mathbf{S})$, $\mu_0(\mathbf{X},\mathbf{S})$, $e(\mathbf{X})$, $m_1(\mathbf{X})$, $m_0(\mathbf{X})$ to be quite general. We propose and evaluate two different versions of this proposed estimator: one where we estimate these components using the Super Learner (Van der Laan *and others*, 2007), which finds an optimal combination of a set of candidate models or learners (denoted "DR-SL") and another that uses the relaxed lasso (Meinshausen, 2007) (denoted "DR-lasso"). In addition, our proposed estimators use a sample-splitting scheme (Chernozhukov *and others*, 2017), which we describe in detail in Appendix C of the Supplementary material available at *Biostatistics* online, to avoid placing restrictive conditions on the estimation of the nuisance functions.

Remark 1 While $\Delta_{\mathbf{S}}$ depends on the counterfactual quantities $\mathbf{S}^{(a)}$ and is thus not the same target treatment effect estimated by Chernozhukov *and others* (2017), because it is identified by the same observed data

quantities in (4.3), we can use the same machinery used in Chernozhukov *and others* (2017), though requiring the stronger assumption (A.1).

4.3. *Inference*

Under very general conditions, including the standard causal assumptions identified in Section 3.1, $\widehat{\Delta}_{\mathbf{S}}$ will converge at the parametric $n^{-\frac{1}{2}}$ rate and be asymptotically normal (Chernozhukov *and others*, 2017; Farrell, 2015) so long as

$$\mathbb{E}\left[\{\widehat{\mu}_a(\mathbf{X},\mathbf{S}) - \mu_a(\mathbf{X},\mathbf{S})\}^2\right] \times \mathbb{E}\left[\{\widehat{\pi}_a(\mathbf{X},\mathbf{S}) - \pi_a(\mathbf{X},\mathbf{S})\}^2\right] = o_p(n^{-\frac{1}{2}}), \tag{4.4}$$

with similar results holding for $\widehat{\Delta}$ if

$$\mathbb{E}\left[\{\widehat{m}_a(\mathbf{X}) - m_a(\mathbf{X})\}^2\right] \times \mathbb{E}\left[\{\widehat{e}_a(\mathbf{X}) - e_a(\mathbf{X})\}^2\right] = o_p(n^{-\frac{1}{2}}). \tag{4.5}$$

As these estimators are built on the efficient influence functions, they are also semiparametrically efficient. Furthermore, as we show in Appendix C of the Supplementary material available at *Biostatistics* online, as long as the sample splits are the same for $\widehat{\Delta}$ and $\widehat{\Delta}_{\mathbf{S}}$, we will have $n^{\frac{1}{2}}(\widehat{R}_{\mathbf{S}} - R_{\mathbf{S}}) \longrightarrow N(0,\sigma^2)$ if both (4.4) and (4.5) hold. Specifically, we have that

$$\sigma^2 = \Delta^{-2}\mathbb{E}\left\{\phi_1(\mathbf{O},\Delta,e,m_0,m_1)^2\right\} + \Delta_{\mathbf{S}}^2\Delta^{-4}\mathbb{E}\left\{\phi_2(\mathbf{O},\Delta_{\mathbf{S}},\pi,\mu_0,\mu_1)^2\right\} - \tag{4.6}$$

$$2\Delta_{\mathbf{S}}\Delta^{-3}\mathbb{E}\left\{\phi_1(\mathbf{O},\Delta,e,m_0,m_1)\phi_2(\mathbf{O},\Delta_{\mathbf{S}},\pi,\mu_0,\mu_1)\right\}, \tag{4.7}$$

where

$$\phi_1(\mathbf{O};\Delta,e,m_0,m_1) = \frac{AY - \{A - e_1(\mathbf{X})\}m_1(\mathbf{X})}{e_1(\mathbf{X})} - \frac{(1-A)Y - \{1-A-e_0(\mathbf{X})\}m_0(\mathbf{X})}{e_0(\mathbf{X})} - \Delta,$$

$$\phi_2(\mathbf{O};\Delta_{\mathbf{S}},\pi,\mu_0,\mu_1) = \frac{AY - \{A - \pi_1(\mathbf{X},\mathbf{S})\}\mu_1(\mathbf{X},\mathbf{S})}{\pi_1(\mathbf{X},\mathbf{S})} - \frac{(1-A)Y - \{1-A-\pi_0(\mathbf{X},\mathbf{S})\}\mu_0(\mathbf{X},\mathbf{S})}{\pi_0(\mathbf{X},\mathbf{S})} - \Delta_{\mathbf{S}}.$$

REMARK 2 When the dimension of $\mathbf{S}$ is not high, (4.4) and (4.5) are not in general restrictive, as many parametric and nonparametric methods are able to achieve the slow rates required of the estimators under certain conditions, including the lasso (Tibshirani, 1996), random forests (Wager and Walther, 2015), and deep neural networks (Farrell *and others*, 2021), when used with a sample-splitting scheme. Ensemble methods such as the Super Learner (Van der Laan *and others*, 2007) may also be used to combine these and other methods to obtain good performance if any of the methods achieves the required rates of consistency. As described in the prior section, our proposed approach combines both ensemble estimation of nuisance functions and sample splitting; we demonstrate the good performance of this approach in finite samples in Section 5.

In higher dimensions, the convergence in (4.4) and (4.5) is more difficult to ensure without further restrictions. If only $\mathbf{S}$ is high-dimensional, then a typical approach is to assume that μ_a may be specified as a sparse linear model $\mu_a(\mathbf{x},\mathbf{s}) = \beta_0 + \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_2^{\mathsf{T}}\mathbf{s}$ and π may be specified as a sparse logistic regression model $\pi_1(\mathbf{x},\mathbf{s}) = g(\delta_0 + \boldsymbol{\delta}_1^{\mathsf{T}}\mathbf{x} + \boldsymbol{\delta}_2^{\mathsf{T}}\mathbf{s})$ for $g(x) = e^x/(1+e^x)$, and with $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2$ sparse enough to ensure (4.4). Because of the fact that $e_a(\mathbf{x}) = P(A = a|\mathbf{X} = \mathbf{x})$ is related to $\pi_a(\mathbf{x},\mathbf{s}) = P(A = a|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s})$, restricting $\pi_1(\mathbf{x},\mathbf{s})$ to this class of models is unproblematic so long as $\mathbf{X}$ is low-dimensional and $e_1(\mathbf{x})$ may be estimated nonparametrically (i.e., without being restricted to the class of sparse linear models). However, if $\mathbf{X}$ is also high-dimensional, sparse logistic models for $e_1(\mathbf{x})$ and $\pi_1(\mathbf{x},\mathbf{s})$ may not in general be

compatible with one another because of the well-known noncollapsibility of logistic regression (Guo and Geng, 1995), unless $\mathbf{S} \perp\!\!\!\perp A|\mathbf{X}$ (which would imply $R_{\mathbf{S}} = 0$) or $\mathbf{S} \perp\!\!\!\perp \mathbf{X}|A$ (which would imply that $\mathbf{X}$ does not confound the relationship between A and $\mathbf{S}$). In these cases, the approach of Tan (2020) may be used to estimate the nuisance functions, in which case only one of $e_1(\mathbf{x})$ and $\pi_1(\mathbf{x}, \mathbf{s})$ are required to be a correctly specified logistic regression. However, this approach did not appear to perform well in our simulations in Section 5. Despite these theoretical considerations, our simulations suggest that an ensemble approach using the Super Learner may outperform approaches based on sparse linear models, as in Tan (2020) or in the version of our estimator that uses only lasso regression. More theoretical development in this area may be warranted.

5. Simulations

5.1. *Simulation overview*

Our proposed estimator based on the Super Learner ("DR-SL") is implemented using the `SuperLearner` package (Polley *and others*, 2021) in `R`. The candidate learners we included were the lasso, ridge regression, ordinary least squares, support vector machines, and random forests for μ_1, μ_0, m_1, m_0 and the lasso, logistic regression, linear discriminant analysis, quadratic discriminant analysis, support vector machines, and random forests for π and e. Our second estimator using the relaxed lasso ("DR-lasso") is implemented using the `glmnet` package (Friedman *and others*, 2010), to estimate all needed functions. For all simulations, we truncated estimates of the propensity and surrogate scores so that $0.01 \leq \widehat{\pi}(\mathbf{X}_i, \mathbf{S}_i), \widehat{e}(\mathbf{X}_i) \leq 0.99$ for all i. We used the default cross-validation procedure to select the tuning parameters for the Super Learner. For the DR-lasso estimator, we used the default crossvalidation to select the tuning parameters. It is possible that performance could be improved by tweaking the tuning parameters for the candidate learners.

It is our understanding that there are no currently available methods to robustly estimate the PTE of a high-dimensional surrogate. However, since there are available methods to measure the strength of high-dimensional mediators, we compare our proposed approach to these available methods. While the goals of mediation analysis and surrogate markers analysis are different and the necessary assumptions differ, this allows us to offer some reasonable comparison to methods that are currently available, rather than no comparison at all. Thus, to fairly compare, we set up the majority of our simulations such that $\mathbf{S}$ is a mediator in that it lies on the causal pathway between A and Y. While mediation methods are attempting to estimate a distinct quantity from $R_{\mathbf{S}}$, in three of the four following simulation settings, the estimation of the "proportion mediated" (which is used in mediation) and $R_{\mathbf{S}}$ are the same.

In our simulations, we compare our proposed approach to: high-dimensional mediation analysis (HIMA, Zhang *and others*, 2016) as implemented in the `HIMA` package (Zheng *and others*, 2018); Bayesian mediation analysis (BAMA, Song *and others*, 2020) as implemented in the `bama` package (Rix *and others*, 2021); and high-dimensional linear mediation analysis (Zhou *and others*, 2020) as implemented in the `freebird` package. All three of these methods propose $p + 1$ linear models: p models for the surrogates/mediators—$S_{ij} = \alpha_{0j} + \alpha_j A_i + \sum_{k=1}^{q} \boldsymbol{\gamma}_{jk}^* X_{ik} + e_{ij}, j = 1, ..., p; i = 1, ..., n$—and one outcome model—$Y_i = \beta_0 + \Delta_S A_i + \sum_{j=1}^{p} \beta_j S_{ij} + \sum_{k=1}^{q} \gamma_k X_{ik} + \epsilon_i$—though the implementation of `freebird` does not allow for the inclusion of covariates X_{ik}. Using these models, the overall treatment effect can be identified as $\Delta = \Delta_S + \sum_{j=1}^{p} \alpha_j \beta_j$, and as above the PTE (or proportion mediated) may be identified as $R_{\mathbf{S}} = 1 - \Delta_S/\Delta$.

We computed 95% confidence intervals (CIs) for $R_{\mathbf{S}}$ as $\widehat{R}_{\mathbf{S}} \pm 1.96\widehat{\sigma}$, where $\widehat{\sigma}$ is an estimate of (4.6). We computed 95% credible intervals for the BAMA estimator from the 2.5% and 97.5% quantiles of the posterior distribution of $1 - \Delta_S / \left(\Delta_S + \sum_{j=1}^{p} \alpha_j \beta_j\right)$. CIs were not computed for the HIMA and freebird implementations because they are not available.

5.2. *Simulation setup*

We constructed four sets of simulations for a total of 18 simulation settings to assess the performance of our proposed approach in both low- and high-dimensional settings.

For the first set of simulations, the data-generating mechanism for $\mathbf{S}^{(a)}$ was linear in $\mathbf{X}$, the data-generating mechanism for $Y^{(a)}$ was linear in $\mathbf{X}$ and $\mathbf{S}^{(a)}$, and the propensity score was linear on the log-odds scale. Given this data-generating mechanism, we would expect that all methods (proposed and comparisons) should perform reasonably well. We set the dimension of $\mathbf{S}$ (p) and of $\mathbf{X}$ (q) to be 100. We let $X_{ij} \sim N(0,1), A_i \sim \text{Bernoulli}\{\pi_1(\mathbf{X}_i)\}, \pi_1(\mathbf{X}_i) = \text{logit}(\boldsymbol{\gamma}^\mathsf{T}\mathbf{X}_i)$, where $\boldsymbol{\gamma} \sim N(0,1)$. The surrogates were generated as $\mathbf{S}_i^{(a)} = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_a^\mathsf{T}\mathbf{X}_i + \mathbf{e}_i$, where $\alpha_{11} = 0.75, \alpha_{12} = 0.25, \alpha_{01} = \alpha_{02} = 0, \alpha_{1j} \sim U(0,1), \alpha_{0j} \sim U(-0.5, 0.5), j = 3, \ldots p$. And only five of the covariates were important for determining the surrogate: $(\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}) = (-1, -0.5, 0, 0.5, 1), (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (-2, -1.5, -1, -0.5, 0), \beta_{aj} = 0, a = 0, 1, j > 5$. The outcome counterfactuals were given by $Y_i^{(a)} = a\Delta_S + \sum_{j=1}^{25} X_{ij} + S_{i1}^{(a)} + S_{i2}^{(a)} + \epsilon_i$, suggesting that only the first two surrogates and 25 of the covariates were important for determining the outcome. The errors e_{ij} and ϵ_i were $N(0, \sigma^2)$. We let the sample size vary between $n = 100$ and $n = 500$ and the level of noise between $\sigma = 0.1$ and $\sigma = 0.5$—for a total of four settings—and we set $R_\mathbf{S} = 0.5$.

In the second set of simulations (six more settings), the data-generating mechanism was less linear, including interactions between covariates in both the propensity score and the model for $\mathbf{S}$, though the outcome model was still a simple linear combination of $\mathbf{X}$ and $\mathbf{S}$. In the third set of simulations (two more settings), the dimension of the surrogates ($p = 1,000$) was much larger than the sample size ($n = 50, 200$), and correct specification of the outcome functions also included nonlinear terms. These simulations should mimic what may happen in practice since all models are typically subject to some amount of mis-specification. We expected the nonlinearity to induce bias in the competing methods (which require linear models to hold). In the fourth set of simulations (six more settings), we specified $\mathbf{S}$ so that it was not a mediator but was instead downstream of a true mediator. We performed 1000 replications for each simulation setting. We give full details and results for these simulations in Appendix E of the Supplementary material available at *Biostatistics* online.

5.3. *Results*

As expected, the first data-generating mechanism was well approximated using linear models. When the sample size was large ($n = 500$) and $\sigma = 0.5$, all methods performed well, with a relatively small bias (see Figure 2). The median $\widehat{R}_\mathbf{S}$ estimates were 0.46 (BAMA), 0.48 (freebird), 0.49 (HIMA), 0.48 (DR-SL), and 0.49 (DR-lasso), all of which compare favorably to the true $R_\mathbf{S}$ value of 0.5. The distribution of estimates for the proposed approaches was a bit tighter than for the competing methods. The empirical 2.5% and 97.5% quantiles were 0.32 and 0.60 for DR-SL, 0.35 and 0.59 for DR-lasso, 0.15 and 0.67 for HIMA, 0.11 and 0.63 for BAMA, and -0.85 and 0.71 for freebird. Results were similar for the sample size with less noise ($\sigma = 0.1$), except for the freebird approach which began to estimate the PTE to be exactly 1 in almost all simulations: the empirical 2.5% and 97.5% quantiles of the $\widehat{R}_\mathbf{S}$ distribution for freebird were exactly 1. At the lower sample size ($n = 100$), HIMA did not run, while BAMA estimates tended to be near 0 and freebird estimates were again clustered tightly around 1. Our proposed approaches had median $\widehat{R}_\mathbf{S}$ values of 0.45 (DR-SL) and 0.56 (DR-lasso), though with more variability than when $n = 500$. The median absolute deviation ($|\widehat{R}_\mathbf{S} - R_\mathbf{S}|$) was smaller for both DR-SL and DR-lasso than any of the competing approaches for all settings.

CI coverage for the proposed estimators tended to be quite good for DR-SL in all settings: coverage was 95% ($n = 100, \sigma = 0.1$), 96% ($n = 100, \sigma = 0.5$), 94% ($n = 500, \sigma = 0.1$), and 97% ($n = 500, \sigma = 0.5$). Coverage for the DR-lasso estimator was worse in general: coverage was 88% ($n = 100, \sigma = 0.1$), 82% ($n = 100, \sigma = 0.5$), 89% ($n = 500, \sigma = 0.1$), and 96% ($n = 500, \sigma = 0.5$) However, BAMA CIs also

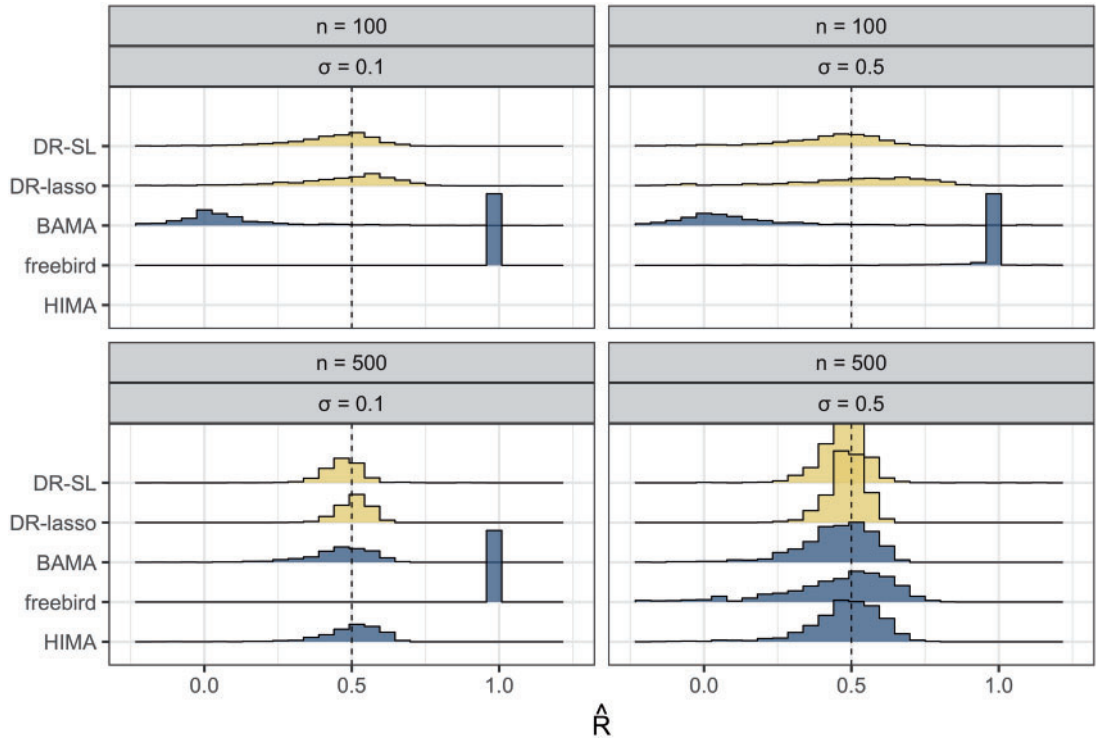

Fig. 2. Distribution of estimates of $\widehat{R}_S$ in the first data-generating mechanism. Lighter shaded regions represent the distribution of the proposed estimators ("DR-SL" and "DR-lasso"); darker shaded regions represent the distribution of the comparison estimators ("BAMA," "freebird," and "HIMA"). The true value of R_S is given as a vertical dotted line at 0.5. At the lower sample size ($n = 100$), HIMA did not run and so is not shown.

did not uniformly obtain nominal coverage: 96% ($n = 100, \sigma = 0.1$), 99% ($n = 100, \sigma = 0.5$), 92% ($n = 100, \sigma = 0.1$), and 85% ($n = 500, \sigma = 0.5$). Even when BAMA CIs had nominal coverage, they were more than four times larger than the CIs for the proposed estimators: for example, BAMA CI half-lengths were 1.40 ($n = 100, \sigma = 0.1$) and 1.61 ($n = 100, \sigma = 0.5$), while the corresponding half-lengths were 0.25 and 0.30 for DR-SL and 0.28 and 0.44 for DR-lasso.

The proposed estimators also outperformed the competing methods in the additional simulation settings where models were subject to mis-specification and the dimensionality of the surrogates was much larger than the sample size. See Appendix E of the Supplementary material available at *Biostatistics* online for complete results.

6. Ebola immune response application

The concentration of binding antibodies is often used as the primary outcome of interest in studies of Ebola vaccine efficacy, being itself a surrogate of vaccine efficacy as measured by the effect on the incidence of infections (Roozendaal *and others*, 2020). Because gene expression is the means by which DNA is turned into RNA and eventually proteins, it is associated with cellular function. Thus, the establishment of the humoral immune response may be captured by changes in gene expression as suggested by early works on systems vaccinology (Li *and others*, 2014). Furthermore, gene expression changes may occur days or even weeks before traditional measures of immune function (Rechtien *and others*, 2017). If gene

expression can act as a surrogate for immune response, then it could possibly be used to shorten vaccine trials or to quickly measure the effect of vaccination in a population. Finally, genome-wide expression data offer the opportunity at looking at very different ways of influencing the response to intervention which constitutes potential surrogate markers. In this study, we sought to use observational data on long-term Ebola survivors and healthy controls to shed light on the possibility of gene expression's use as a surrogate for antibody response to Ebola virus. This aim is inspired by the study of potential surrogates of protection among Ebola disease survivors (Sullivan *and others*, 2009).

In total, 26 Ebola survivors of the 2013–2016 Ebola outbreak in West Africa were recruited from the Postebogui cohort (Etard *and others*, 2017) as well as 33 healthy donors as described in Wiedemann *and others* (2020), each of whom had an expression for 29 624 genes quantified from whole blood RNA-seq (freely available from the Gene Expression Omnibus repository with accession code GSE143549) as well as the concentration of specific Ebola binding antibodies measured. Clearly, this is a setting where the number of potential surrogate markers (the genes) is substantially larger than the sample size. Given the superior performance of the DR-SL in our simulation study, and the high-dimensional setting, we focus only on using the DR-SL estimator to quantify how much of the overall difference in humoral immune response between survivors and healthy donors could be captured by the measured gene expression data. We used the same candidate learners as specified in Section 5. Propensity and surrogate scores ($\widehat{e}$ and $\widehat{\pi}$) were truncated at 0.05 and 0.95 to prevent instability due to extreme weights, and $\mathbf{X}$ included age and sex.

Ebola survivors were estimated to have a much higher abundance of Ebola-specific antibodies ($\widehat{\Delta} = 3,998$, SE $= 851.5$). Using DR-SL, the residual treatment effect, $\Delta_{\mathbf{S}}$, was estimated as $\widehat{\Delta}_{\mathbf{S}} = 3,242$ with SE $= 727.8$, and the proportion of the difference explained by gene expression was estimated as $\widehat{R}_{\mathbf{S}} = 0.1890$, with a SE of 0.07923. Thus, a large part of the humoral immune response cannot be explained by the differences in gene expression. Of note, we assumed no unmeasured confounding factors although it cannot be guaranteed in such a real-life context where survivors and healthy volunteers are two selected populations. If unmeasured confounding inflated both $\widehat{\Delta}$ and $\widehat{\Delta}_{\mathbf{S}}$ roughly equally, this could have the effect of artificially deflating $\widehat{R}_{\mathbf{S}}$. Another explanation for the low estimated $\widehat{R}_{\mathbf{S}}$ could be that, while gene expression measured shortly after infection may potentially be a good surrogate, measuring it long after infection (as in this study), does not capture the treatment effect as well. Measurement error in $\mathbf{S}$ could also deflate the PTE. Importantly, one other potential violation of our assumptions is that about one-third of the observations have truncated surrogate scores—$\widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i) < 0.05$ or $\widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i) > 0.95$. This suggests that positivity might be (nearly) violated. We discuss this further in Appendix C of the Supplementary material available at *Biostatistics* online along with a potential solution.

7. DISCUSSION

We have proposed a very general approach to evaluating surrogate markers which can be applied in randomized experiments or in observational studies and can be used regardless of the dimensionality of the surrogates. Our approach is robust in that we have defined the PTE of the surrogates without reference to any models, and we have shown how machine learning approaches like Super Learner may be used to very flexibly estimate nuisance functions. Our simulation results suggest that our Super Learner estimator (DR-SL) outperforms competing methods even when the underlying data-generating mechanism is linear and still gives reasonable results even when high-dimensional linear models are mis-specified.

Our approach here is intimately tied to previous approaches for evaluating surrogate markers. The approaches in Agniel and Parast (2021) and in Parast *and others* (2016) can be seen to be similar to a version of our approach where the average treatment effect among controls is estimated under further assumptions. In Agniel and Parast (2021), they further require strict randomization of treatment, and they assume that $\mathbf{S}$ is a realization of a smooth continuous function. Parast *and others* (2016) make similar assumptions but take S to be a scalar surrogate. They estimate a version of (4.2) among controls:

$\Delta_{\mathbf{S}} = \mathbb{E}\{\mu_1(\mathbf{S}) - \mu_0(\mathbf{S})|A = 0\}$ using kernel smoothing and taking advantage of the fact that treatment is randomized. Their estimates have the form $\widehat{\Delta}_{\mathbf{S}} = n_0^{-1} \sum_{A_i=0} \{\widehat{\mu}_1(\mathbf{S}_i) - Y_i\}$, where $n_0 = \sum_{i=1}^n I\{A_i = 0\}$ and $\widehat{\mu}_1(\cdot)$ is estimated via kernel smoothing (possibly after dimension reduction). Our approach here could easily be adapted to estimate a similar quantity by using methods for doubly robust estimation of the average treatment effect on the treated (Shu and Tan, 2018; Moodie *and others*, 2018; Chernozhukov *and others*, 2017) by, for example, $\widehat{\Delta}_{\mathbf{S}} = n_0^{-1} \sum_{i=1}^n \left\{A_i \frac{1-\widehat{\pi}(\mathbf{S}_i)}{\widehat{\pi}(\mathbf{S}_i)} - (1 - A_i)\right\} \{\widehat{\mu}_1(\mathbf{S}_i) - Y_i\}$, where $\widehat{\pi}(\mathbf{s}) = \widehat{\mathbb{P}}(A_i|\mathbf{S}_i = \mathbf{s})$ may be estimated using a model similar to the one used for $\widehat{\mu}_1(\mathbf{s})$. Furthermore, our approach could be used to facilitate the use of machine learning methods in the estimation of $\widehat{\mu}_1(\cdot), \widehat{\pi}(\cdot)$, to include covariates to control for confounding, or to simplify or strengthen asymptotic results. For example, results for the kernel estimator used in Agniel and Parast (2021) obtain rates of convergence of $n^{-\frac{1}{2}}$ only under very limited technical conditions, but using sample-splitting parametric rates of convergence could be obtained under very general conditions.

8. Software

We include software to implement our proposed methods in the R package `crossurr` available at `github.com/denisagniel/crossurr`.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Yves Lévy and all of the investigators at the Vaccine Research Institute for providing the Ebola survivor data. We thank the Postebogui team for their daily work and the survivors. This work was supported by grant R01 DK118354 from the National Institute of Diabetes and Digestive and Kidney Diseases as well as by the Investissements d'Avenir program managed by the ANR under reference ANR-10-LABX-77-01. *Conflict of Interest*: None declared.

References

Agniel, D. and Parast, L. (2021). Evaluation of longitudinal surrogate markers. *Biometrics* **77**, 477–489.

Alonso, A., Molenberghs, G. *and others*. (2004). Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics* **60**, 724–728.

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M. and Burzykowski, T. (2016). An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics* **72**, 669–677.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Caveney, E. J. and Cohen, O. J. (2011). Diabetes and biomarkers. *Journal of Diabetes Science and Technology* **5**, 192–197.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review* **107**, 261–265.

Choi, S. H., Kim, T. H., Lim, S., Park, K. S., Jang, H. C. and Cho, N. H. (2011). Hemoglobin A1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction: a 6-year community-based prospective study. *Diabetes Care* **34**, 944–949.

Conlon, A., Taylor, J., Li, Y., Diaz-Ordaz, K. and Elliott, M. (2017). Links between causal effects and causal association for surrogacy evaluation in a Gaussian setting. *Statistics in Medicine* **36**, 4243–4265.

ETARD, J.-F., SOW, M. S., LEROY, S., TOURÉ, A., TAVERNE, B., KEITA, A. K., MSELLATI, P., BAIZE, S., RAOUL, H., IZARD, S. AND KPAMOU, C. (2017). Multidisciplinary assessment of post-Ebola sequelae in Guinea (Postebogui): an observational cohort study. *The Lancet Infectious Diseases* **17**, 545–552.

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189**, 1–23.

FARRELL, M. H., LIANG, T. AND MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89**, 181–213.

FREEDMAN, L. S., GRAUBARD, B. I. AND SCHATZKIN, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

GILBERT, P. B. AND HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154.

GUO, J. AND GENG, Z. (1995). Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 263–267.

JOFFE, M. M. AND GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**, 530–538.

KARIM, S. S. A. (2021). Vaccines and SARS-CoV-2 variants: the urgent need for a correlate of protection. *The Lancet* **397**, 1263–1264.

LI, S., ROUPHAEL, N. *and others*. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology* **15**, 195–204.

MEINSHAUSEN, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52**, 374–393.

MOODIE, E. E. M., SAARELA, O. AND STEPHENS, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat* **7**, e205.

OBIRIKORANG, C., QUAYE, L. AND ACHEAMPONG, I. (2012). Total lymphocyte count as a surrogate marker for CD4 count in resource-limited settings. *BMC Infectious Diseases* **12**, 1–5.

PARAST, L., CAI, T. AND TIAN, L. (2019). Using a surrogate marker for early testing of a treatment effect. *Biometrics* **75**, 1253–1263.

PARAST, L., MCDERMOTT, M. M. AND TIAN, L. (2016). Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in Medicine* **35**, 1637–1653.

PARAST, L., TIAN, L. AND CAI, T. (2020). Assessing the value of a censored surrogate outcome. *Lifetime Data Analysis* **26**, 245–265.

PLOTKIN, S. A. AND GILBERT, P. B. (2012). Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases* **54**, 1615–1617.

POLLEY, E., LEDELL, E., KENNEDY, C., LENDLE, S. AND VAN DER LAAN, M. (2021). SuperLearner: Super Learner Prediction. R package version 2.0-28. https://CRAN.R-project.org/package=SuperLearner.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

PRICE, B. L., GILBERT, P. B. AND VAN DER LAAN, M. J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics* **74**, 1271–1281.

RECHTIEN, A., RICHERT, L. *and others*. (2017). Systems vaccinology identifies an early innate immune signature as a correlate of antibody responses to the Ebola vaccine rVSV-ZEBOV. *Cell Reports* **20**, 2251–2261.

RIX, A., KLEINSASSER, M. AND SONG, Y. (2021). *BAMA: High Dimensional Bayesian Mediation Analysis*. R package version 1.2. https://CRAN.R-project.org/package=bama

ROOZENDAAL, R., HENDRIKS, J. *and others*. (2020). Nonhuman primate to human immunobridging to infer the protective effect of an Ebola virus vaccine candidate. *NPJ Vaccines* **5**, 1–11.

SHU, H. AND TAN, Z. (2018). Improved estimation of average treatment effects on the treated: local efficiency, double robustness, and beyond. *arXiv preprint arXiv:1808.01408*.

SMALL, G. W. (2006). Diagnostic issues in dementia: neuroimaging as a surrogate marker of disease. *Journal of Geriatric Psychiatry and Neurology* **19**, 180–185.

SONG, Y., ZHOU, X., ZHANG, M., ZHAO, W., LIU, Y., KARDIA, S. L., ROUX, A.V.D., NEEDHAM, B.L., SMITH, J. A. AND MUKHERJEE, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics* **76**, 700–710.

SULLIVAN, N. J., MARTIN, J. E., GRAHAM, B. S. AND NABEL, G. J. (2009). Correlates of protective immunity for Ebola vaccines: implications for regulatory approval by the animal rule. *Nature Reviews Microbiology* **7**, 393–400.

TAN, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics* **48**, 811–837.

TAYLOR, J. M. G., WANG, Y. AND THIÉBAUT, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**, 1102–1111.

TEIXEIRA, A. L., DINIZ, B. S., CAMPOS, A. C., MIRANDA, A. S., ROCHA, N. P., TALIB, L. L., GATTAZ, W. F. AND FORLENZA, O. V. (2013). Decreased levels of circulating adiponectin in mild cognitive impairment and Alzheimer's disease. *Neuromolecular medicine* **15**, 115–121.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288.

VAN DER LAAN, M. J., POLLEY, E. C., AND HUBBARD, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology* **6**, 1–23.

VANDERWEELE, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics* **69**, 561–565.

WAGER, S. AND WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

WANG, S., MCCORMICK, T. H., AND LEEK, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences* **117**, 30266–30275.

WANG, Y. AND TAYLOR, J. M. G. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58**, 803–812.

WIEDEMANN, A., FOUCAT, E. *and others*. (2020). Long-lasting severe immune dysfunction in Ebola virus disease survivors. *Nature Communications* **11**, 1–11.

ZHANG, H., ZHENG, Y. *and others*. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32**, 3150–3154.

ZHENG, Y., ZHANG, H., HOU, L. AND LIU, L. (2018). *HIMA: High-Dimensional Mediation Analysis*. R package version 1.0.7.

ZHOU, R. R., WANG, L. AND ZHAO, S. D. (2020). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika* **107**, 573–589.

ZHOU, R. R., ZHAO, S. D. AND PARAST, L. (2022). Estimation of the proportion of treatment effect explained by a high-dimensional surrogate. *Statistics in Medicine* **41**, 2227–2246.

[*Received December 30, 2021; revised May 16, 2022; accepted for publication June 3, 2022*]