Received: 31 March 2018 | Accepted: 1 May 2020

DOI: 10.1111/biom.13298

BIOMETRIC METHODOLOGY

Biometrics WILEY

Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data

David Cheng[1] | **Ashwin N. Ananthakrishnan**[2] | **Tianxi Cai**[3]

[1]VA Boston Healthcare System, Boston, Massachusetts

[2]Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

Correspondence
Tianxi Cai, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115.
Email: tcai@hsph.harvard.edu

Funding information
National Institutes of Health, Grant/Award Numbers: T32CA009337, R21CA242940, R01HL089778

Abstract

We consider the problem of estimating the average treatment effect (ATE) in a semi-supervised learning setting, where a very small proportion of the entire set of observations are labeled with the true outcome but features predictive of the outcome are available among all observations. This problem arises, for example, when estimating treatment effects in electronic health records (EHR) data because gold-standard outcomes are often not directly observable from the records but are observed for a limited number of patients through small-scale manual chart review. We develop an imputation-based approach for estimating the ATE that is robust to misspecification of the imputation model. This effectively allows information from the predictive features to be safely leveraged to improve efficiency in estimating the ATE. The estimator is additionally doubly-robust in that it is consistent under correct specification of either an initial propensity score model or a baseline outcome model. It is also locally semiparametric efficient under an ideal semi-supervised model where the distribution of the unlabeled data is known. Simulations exhibit the efficiency and robustness of the proposed method compared to existing approaches in finite samples. We illustrate the method by comparing rates of treatment response to two biologic agents for treatment inflammatory bowel disease using EHR data from Partners' Healthcare.

KEYWORDS

causal inference, double-robustness, missing data, semiparametric efficiency, semi-supervised learning, surrogate outcomes

1 | INTRODUCTION

There is often interest in estimating the average treatment effect (ATE) of a binary treatment T on an outcome Y that is observed among a very limited subset of observations but can be approximated by surrogate variables $\mathbf{W}$ available among all observations. As a motivating example, we consider comparing the outcomes of two treatments in electronic health record (EHR) data, where Y can be a clinical outcome of interest not directly encoded in patients' medical records and $\mathbf{W}$ are post-treatment features that can be automatically extracted from the charts, such as the receipt of billing or procedure codes and mentions of selected terms from physicians' notes. Although different study designs are possible, we assume for phenotyping purposes Y is collected for a small *random* subset of all patients, which constitute the labeled data $\mathscr{L}$, through manual chart review. It may not be possible to comprehensively label the data because chart review is a costly and time-consuming process.

A common strategy for analyzing such data is to use the surrogates $\mathbf{W}$ to approximate Y by some imputed outcome $Y^{\dagger} = g(\mathbf{W})$. The imputation can be based on heuristic rules determined by domain-specific knowledge (eg, presence of certain set of diagnostic codes) or an imputation model that

predicts Y given $\mathbf{W}$ trained using the labeled data $\mathscr{L}$, which includes observations of both $\mathbf{W}$ and Y (Ananthakrishnan *et al.*, 2016). However, for complex outcomes it may be difficult to obtain an accurate imputation $Y^{\dagger}$ because $\mathbf{W}$ have limited predictive power or the functional form of the imputation model may be difficult to correctly specify. When the imputation quality is inadequate, it is often unclear whether using the inaccurate imputations $Y^{\dagger}$ in subsequent analyses can lead to biased estimates of the ATE on the *actual* outcome Y.

Previously, related methods have been developed in the surrogate outcomes literature to leverage both the labeled data $\mathscr{L}$ and the unlabeled data $\mathscr{U}$, which includes all variables in $\mathscr{L}$ except for Y, for estimating regression parameters (Pepe, 1992) and solutions to estimating equations (Chen *et al.*, 2003). But these methods tend to assume a univariate surrogate with low-dimensional baseline covariates $\mathbf{X}$. Alternatively the problem can be viewed as estimating the mean of a longitudinal outcome subject to monotone missingness, where $\mathbf{W}$ is an outcome at an initial time point and Y is the final outcome. In this context, semiparametric efficiency theory has been developed to identify efficient estimators under various semiparametric models (Robins *et al.*, 1994; Rotnitzky *et al.*, 1998), which has lead to development of doubly-robust (DR) augmented inverse probability weighting (AIPW) estimators in different problems (Davidian *et al.*, 2005; Williamson *et al.*, 2012; Zhang *et al.*, 2016). In particular, Davidian *et al.* (2005) develop such an efficient estimator for estimating the effect of a randomized treatment where the final outcome is subject to missingness but intermediate outcomes are collected for all patients. With some minor modification this estimator could also leverage the unlabeled data $\mathscr{U}$ to aid estimation of the ATE, and we consider it as a reference method in the simulations. However, this method was developed under a data model commonly employed in missing data problems with independent and identically (iid) observations and a probability of missingness that is bounded away from 0. The data model we consider here differs in that: (1) we assume that the number of labeled observations n is fixed by design, and (2) we make a *semi-supervised* assumption that the proportion of labeled data ν_n tends to 0 as $n \to \infty$, to reflect the large size of $\mathscr{U}$ relative to that of $\mathscr{L}$. These features complicate conventional applications of semiparametric efficiency theory, and the efficiency and finite sample performance of existing estimators may not be clear.

In this paper, we propose a semi-supervised (SS) estimator for the ATE based on an imputation followed by IPW. It is doubly-robust and locally semiparametric efficient under an ideal model approximating the data distribution of the semi-supervised setup. The imputations are constructed such that the resulting estimator is robust to misspecification of the imputation model, enabling $\mathscr{U}$ to be safely used to improve the estimation. We further employ a double-index propensity score (Cheng *et al.*, 2019) for additional robustness and possible small-sample efficiency gains. The remainder of the paper is organized as follows. We formalize the SS estimation problem in Sections 2.1 and 2.2 and develop the estimator in Sections 2.3-2.5. A perturbation resampling procedure is proposed in Section 2.6 for inference. Section 3 presents simulations showing the robustness and efficiency of the proposed estimator, and Section 4 applies the method to compare two biologic therapies for treating inflammatory bowel disease (IBD) in EMR data from Partners' Healthcare. We conclude with some remarks in Section 5. Proofs are deferred to the Web Appendices.

2 | METHOD

2.1 | Notations and semi-supervised framework

Let Y denote an outcome, $T \in \{0, 1\}$ a binary treatment, $\mathbf{X}$ a p_x-dimensional vector of pre-treatment baseline covariates, $\mathbf{W}$ a p_w-dimensional vector of post-treatment surrogate variables that are potentially predictive of Y, and $\mathbf{V} = (\mathbf{W}^{\mathsf{T}}, \mathbf{X}^{\mathsf{T}})^{\mathsf{T}}$. For example, in the EHR context, $\mathbf{X}$ may include demographics and prior comorbidities that may confound naive associations between T and Y, while $\mathbf{W}$ may be counts of post-treatment codes or terms. The labeled data consist of n iid observations $\mathscr{L} = \{(Y_i, T_i, \mathbf{V}_i^{\mathsf{T}})^{\mathsf{T}} : i = 1, \ldots, n\}$, while the unlabeled data consist of $N - n$ iid observations without Y, $\mathscr{U} = \{(T_i, \mathbf{V}_i^{\mathsf{T}})^{\mathsf{T}} : i = n+1, \ldots, N\}$, with $\mathscr{U} \perp\!\!\!\perp \mathscr{L}$ and n and N fixed. We assume that n observations were randomly selected for labeling so that Y is essentially missing completely at random (MCAR) from observations in $\mathscr{U}$. In the SS setting, $N \gg n$ so that $\nu_n = n/N \to 0$ as $n \to \infty$. The entire observed data $\mathscr{D} = \mathscr{L} \cup \mathscr{U}$ could thus be framed as $\mathscr{D} = \{(L_i, \widetilde{Y}_i, T_i, \mathbf{V}_i^{\mathsf{T}}) : i = 1, \ldots, N\}$, where L is an indicator of labeling such that $\widetilde{Y} = Y$ when $L = 1$ and $\widetilde{Y}$ is an arbitrary value otherwise with $L \perp\!\!\!\perp (\widetilde{Y}, T, \mathbf{V})$. But unlike traditional missing data frameworks, L is constrained such that $\sum_{i=1}^{N} L_i = n$, where n and N satisfy $\nu_n = n/N \to 0$.

2.2 | Target parameter and leveraging unlabeled data

Let $Y^{(1)}$ and $Y^{(0)}$ denote the counterfactual outcomes had an individual received treatment or control. Based on the observed data $\mathscr{D}$ we want to estimate the ATE:

$$\Delta = \mathbb{E}\{Y^{(1)}\} - \mathbb{E}\{Y^{(0)}\} = \mu_1 - \mu_0. \tag{1}$$

We require the following standard assumptions to identify Δ:

$$Y = TY^{(1)} + (1 - T)Y^{(0)} \tag{2}$$

$$(Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp T \mid \mathbf{X} \tag{3}$$

$$\pi(\mathbf{x}) \in [\epsilon_\pi, 1-\epsilon_\pi] \text{ for some } \epsilon_\pi > 0 \text{ when } f(\mathbf{x}) > 0, \tag{4}$$

where $\pi(\mathbf{x}) = \mathbb{P}(T = 1 \mid \mathbf{X} = \mathbf{x})$ is the PS and $f(\mathbf{x})$ is the joint density for the covariates. In the typical setting where the outcome is fully observed, the ATE can be identified through the g-formula (Robins, 1986) for a point exposure:

$$\Delta = \mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\} = \mathbb{E}\left\{\frac{I(T=1)Y}{\pi(\mathbf{X})} - \frac{I(T=0)Y}{1-\pi(\mathbf{X})}\right\}, \tag{5}$$

where $\mu_k(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, T = k)$ for $k = 0, 1$. This suggests the usual estimators based on averaging the outcome weighted by IPW weights or averaging estimated outcome models. When the outcome is missing but surrogates $\mathbf{W}$ are observed, the more general g-formula for longitudinal studies can be applied to show that

$$\begin{aligned}\Delta &= \mathbb{E}\left[\mathbb{E}\{\xi_1(\mathbf{V}) \mid \mathbf{X}, T=1\} - \mathbb{E}\{\xi_0(\mathbf{V}) \mid \mathbf{X}, T=0\}\right] \\ &= \mathbb{E}\left\{\frac{I(T=1)\xi_1(\mathbf{V})}{\pi(\mathbf{X})} - \frac{I(T=0)\xi_0(\mathbf{V})}{1-\pi(\mathbf{X})}\right\},\end{aligned}$$

where $\xi_k(\mathbf{v}) = \mathbb{E}(\widetilde{Y} \mid \mathbf{V} = \mathbf{v}, T = k, L = 1) = \mathbb{E}(Y \mid \mathbf{V} = \mathbf{v}, T = k)$ for $k = 0, 1$. This decomposition suggests that, if a consistent estimator for $\xi_k(\mathbf{v})$ is available, then Δ can be estimated by first imputing Y through the $\xi_k(\mathbf{v})$ estimator and then applying IPW or outcome regression methods to the imputed outcome. However, obtaining a consistent estimator for $\xi_k(\mathbf{v})$ may not be feasible without strong modeling assumptions due to the potential high dimensionality of $\mathbf{v}$ and complexity of the functional form of $\xi_k(\mathbf{v})$. In the following we show that even with incorrectly specified models for $\xi_k(\mathbf{v})$, it is still possible to leverage $\mathcal{U}$ in estimating Δ without introducing bias from their misspecification.

2.3 | Robust imputations

Let $U_\pi = I(T=1)/\pi(\mathbf{X}) - I(T=0)/\{1-\pi(\mathbf{X})\}$ denote a utility covariate given $\pi(\mathbf{x})$, assumed momentarily to be known. We now argue that inclusion of such a covariate in an imputation model yields imputations that are robust to misspecification of the model. Suppose we postulate a parametric *working model*, possibly misspecified, for $\xi_k(\mathbf{v})$:

$$\xi_T(\mathbf{V}) = g_\xi(\gamma_0 + \boldsymbol{\gamma}_1^\top \mathbf{h}(\mathbf{V}) + \gamma_2 T + \gamma_3 U_\pi) = g_\xi(\boldsymbol{\gamma}^\top \mathbf{Z}_\pi), \tag{6}$$

where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^\top, \gamma_2, \gamma_3)^\top$, $\mathbf{Z}_\pi = (1, \mathbf{V}^\top, T, U_\pi)^\top$, $g_\xi(\cdot)$ is a specified link function, and $\mathbf{h}(\cdot)$ is a vector of fixed basis expansion functions that can incorporate nonlinear effects. Interactions between $\mathbf{h}(\mathbf{V})$ and T could also be included in the specification without difficulty. We estimate $\boldsymbol{\gamma}$ as $\widehat{\boldsymbol{\gamma}}$, the solution to a penalized estimating equation with ridge regularization:

$$n^{-1}\sum_{i=1}^{n} \mathbf{Z}_{\pi,i}\left\{Y_i - g_\xi(\boldsymbol{\gamma}^\top \mathbf{Z}_{\pi,i})\right\} - \lambda_n \boldsymbol{\gamma}_\circ = \mathbf{0}, \tag{7}$$

where $\boldsymbol{\gamma}_\circ = (0, \boldsymbol{\gamma}_{\{-1\}}^\top)^\top$ with $\boldsymbol{\gamma}_{\{-1\}}$ being the subvector of $\boldsymbol{\gamma}$ that excludes the intercept γ_0 and $\lambda_n = o(n^{-1/2})$ is a tuning parameter, which allows $\widehat{\boldsymbol{\gamma}}$ to have an $n^{-1/2}$ convergence rate. In particular, this class of estimators includes ridge estimators for GLMs based on exponential families with canonical link functions. Ridge regularization is suggested here to improve finite sample performance but is not essential for asymptotic properties discussed below. Other regularization penalties besides the ridge penalty or no regularization can also be used, as long as $\widehat{\boldsymbol{\gamma}}$ maintains an $n^{-1/2}$ convergence rate. Using that Y is MCAR, standard arguments (Web Appendix A) can be used to show that $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \bar{\boldsymbol{\gamma}}$ where $\bar{\boldsymbol{\gamma}}$ solves

$$\mathbb{E}\left[\mathbf{Z}_\pi\left\{Y - g_\xi(\boldsymbol{\gamma}^\top \mathbf{Z}_\pi)\right\}\right] = \mathbf{0},$$

with the expectation being taken *over the entire population* and not restricted only to the $L = 1$ subpopulation. In particular, for $Y^\dagger = g_\xi(\bar{\boldsymbol{\gamma}}^\top \mathbf{Z}_\pi)$, since $\mathbf{Z}_\pi$ includes U_π, this implies

$$\begin{aligned}&\mathbb{E}\left\{\frac{I(T=1)Y}{\pi(\mathbf{X})} - \frac{I(T=0)Y}{1-\pi(\mathbf{X})}\right\} \\ &= \mathbb{E}\left\{\frac{I(T=1)Y^\dagger}{\pi(\mathbf{X})} - \frac{I(T=0)Y^\dagger}{1-\pi(\mathbf{X})}\right\}.\end{aligned} \tag{8}$$

This suggests that a standard IPW estimator based on the imputed outcome $Y^\dagger$ has the *same asymptotic limit* as if the true outcomes were used, even if imputation model (6) is misspecified. Consequently, the surrogate data from $\mathcal{U}$ could be safely used to impute the outcome using estimates of these imputed outcomes $Y^\dagger$.

Augmentation covariates similar to U_π have been previously used, for example, to construct locally semiparametric efficient doubly-robust estimators (Bang and Robins, 2005; van der Laan and Rubin, 2006), construct improved locally efficient doubly-robust estimators with enhanced efficiency under misspecification (Rotnitzky *et al*., 2012), and ensure valid doubly-robust inference (Benkeser *et al*., 2017). Our work follows these previous works in leveraging augmentation covariates to achieve desired statistical properties. In our case, U_π is used to ensure robustness of the final estimator to misspecification of the working imputation model.

In practice, $\pi(\mathbf{x})$ also needs to be estimated, which is typically done through parametric modeling such as logistic regression. When $\pi(\mathbf{x})$ is estimated by an estimator $\widehat{\pi}(\mathbf{x})$, the IPW estimator discussed above will be consistent for Δ if $\widehat{\pi}(\mathbf{x})$ is consistent for $\pi(\mathbf{x})$ but otherwise could be biased if the parametric model for $\pi(\mathbf{x})$ is misspecified. Similar arguments

can be used to construct alternative imputations $Y^\dagger$ that could be substituted for Y in an outcome regression estimator and maintain robustness to misspecification of the imputation model. However, such an approach would then require correct specification of the outcome regression model $\mu_k(\mathbf{x})$ to be consistent for Δ.

2.4 | Doubly-robust IPW based on the double-index PS

We adopt an IPW estimator for the final estimator but weighting with a double-index PS (DiPS), which is an alternative method to estimating the PS $\pi(\mathbf{x})$ that calibrates an initial parametric PS estimate to allow prognostic covariates in $\mathbf{X}$ to inform the PS estimation (Cheng *et al*., 2019). When Y is fully observed, this was previously shown to yield a DR IPW estimator, which is consistent for Δ if a working model for either $\pi(\mathbf{x})$ or $\mu_k(\mathbf{x})$ is correctly specified. Adoption of DiPS in the proposed SS estimator also yields a DR estimator, obviating the need for alternative sets of robust imputations $Y^\dagger$ to accommodate estimation based on correctly specifying models for $\pi(\mathbf{x})$ or $\mu_k(\mathbf{x})$. Plugging in the DiPS is a natural approach to achieve double-robustness in this context, as the rationale for robust imputations from above assumes an IPW estimator is used for the final estimate following imputation. Although augmented IPW (AIPW) estimators (Robins *et al*., 1994) can also be used to achieve double-robustness, it is not immediately clear whether the resulting estimator would be robust to misspecification of imputation model.

In the following, we present the details in estimating DiPS and then define the final IPW estimator. We postulate the following working parametric models for $\pi(\mathbf{x})$ and $\mu_k(\mathbf{x})$:

$$\pi(\mathbf{X}) = g_\pi(\alpha_0 + \boldsymbol{\alpha}_1^\mathsf{T}\mathbf{X}) = \pi(\mathbf{X};\boldsymbol{\alpha}) \tag{9}$$

$$\mu_T(\mathbf{X}) = g_\mu(\beta_0 + \boldsymbol{\beta}_1^\mathsf{T}\mathbf{X} + \beta_2 T) = \mu_T(\mathbf{X};\boldsymbol{\beta}), \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1^\mathsf{T})^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\mathsf{T}, \beta_2)^\mathsf{T}$, and $g_\pi(\cdot)$ and $g_\mu(\cdot)$ are specified link functions. Interactions between $\mathbf{X}$ and T in the baseline outcome model $\mu_k(\mathbf{x};\boldsymbol{\beta})$ can also be accommodated by estimating the DiPS separately by treatment groups (Cheng *et al*., 2019). To allow for a large set of covariates, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated using regularized maximum likelihood estimators, for example, as in $\widehat{\boldsymbol{\alpha}} = \text{argmin}_\alpha\{-N^{-1}\sum_{i=1}^N \ell_\pi(\boldsymbol{\alpha};\mathbf{X}_i,T_i) + p_{\lambda_N}(\boldsymbol{\alpha}_1)\}$ and $\widehat{\boldsymbol{\beta}} = \text{argmin}_\beta\{-n^{-1}\sum_{i=1}^n \ell_\mu(\boldsymbol{\beta};Y_i,\mathbf{X}_i,T_i) + p_{\lambda_n}(\boldsymbol{\beta}_{\{-1\}})\}$, where $\ell_\pi(\boldsymbol{\alpha};\mathbf{X}_i,T_i)$ and $\ell_\mu(\boldsymbol{\beta};Y_i,\mathbf{X}_i,T_i)$ are the log-likelihood contributions for the ith observation, and $p_{\lambda_N}(\cdot)$ and $p_{\lambda_n}(\cdot)$ are penalty functions chosen such that the oracle properties (Fan and Li, 2001) hold, such as the adaptive LASSO (ALASSO) (Zou, 2006). We then calibrate the initial PS estimate $\pi(\mathbf{x};\widehat{\boldsymbol{\alpha}})$ by the kernel smoothing estimator:

$$\widehat{\pi}(\mathbf{x};\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1) = \frac{N^{-1}\sum_{j=1}^N K_h(\widehat{\mathbf{S}}_j - \widehat{\mathbf{s}})I(T_j=1)}{N^{-1}\sum_{j=1}^N K_h(\widehat{\mathbf{S}}_j - \widehat{\mathbf{s}})},$$

where $\widehat{\mathbf{S}}_j = (\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)^\mathsf{T}\mathbf{X}_j$ and $\widehat{\mathbf{s}} = (\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)^\mathsf{T}\mathbf{x}$ are bivariate scores that represent the covariate in the directions of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}_1$, $K_h(\cdot) = h^{-2}K(\cdot/h)$, with $K(\cdot)$ being a bivariate qth order kernel with $q > 2$ and $h = O(N^{-\alpha})$ being a bandwidth for which a suitable choice of α is discussed below. The intution is that smoothing T over $\widehat{\boldsymbol{\alpha}}_1^\mathsf{T}\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}_1^\mathsf{T}\mathbf{X}$ calibrates $\pi(\mathbf{x};\widehat{\boldsymbol{\alpha}})$ closer to the true $\pi(\mathbf{x})$, incorporating variation in covariates in $\mathbf{X}$ that are associated with Y but may have been selected out in the initial PS model due to weak association with T.

Finally, we define the proposed SS_{DR} estimator as $\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0$, where

$$\widehat{\mu}_1 = \left\{\sum_{i=1}^N \frac{I(T_i=1)}{\widehat{\pi}(\mathbf{X}_i;\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)}\right\}^{-1}\left\{\sum_{i=1}^N \frac{I(T_i=1)\widehat{Y}_i^\dagger}{\widehat{\pi}(\mathbf{X}_i;\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)}\right\} \tag{11}$$

$$\text{and } \widehat{\mu}_0 = \left\{\sum_{i=1}^N \frac{I(T_i=0)}{1-\widehat{\pi}(\mathbf{X}_i;\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)}\right\}^{-1} \times \left\{\sum_{i=1}^N \frac{I(T_i=0)\widehat{Y}_i^\dagger}{1-\widehat{\pi}(\mathbf{X}_i;\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\beta}}_1)}\right\}, \tag{12}$$

with $\widehat{Y}_i^\dagger = g_{\widehat{\xi}}(\widehat{\boldsymbol{\gamma}}^\mathsf{T}\mathbf{Z}_{\widehat{\pi},i})$. The final estimator $\widehat{\Delta}$ substitutes the robust imputations $\widehat{Y}^\dagger$, based on the PS estimated by the DiPS, into a standardized IPW estimator weighted also with the DiPS. We next consider the large-sample properties of $\widehat{\Delta}$, starting with its DR property and its influence function expansion.

2.5 | Consistency and asymptotic linearity of $\widehat{\Delta}$

We show in Web Appendix B that, under the causal identification assumptions (2)-(4) and mild regularity conditions, given that $h = O(N^{-\alpha})$ with $\alpha \in (\frac{1-\beta}{2q}, \frac{\beta}{2} \wedge \frac{1}{4})$ and $n = O(N^{1-\beta})$ with $\beta \in (\frac{1}{q+1}, 1)$, $\widehat{\Delta}$ is DR so that

$$\widehat{\Delta} - \Delta = O_p(n^{-1/2}), \tag{13}$$

when either the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ in (9) or the baseline outcome model $\mu_k(\mathbf{x};\boldsymbol{\beta})$ in (10) is correctly specified. As this does not depend on the correct specification of the model for $\xi_k(\mathbf{v})$, this result also verifies that $\widehat{\Delta}$ is robust to misspecification of the imputation model. To characterize the large sample variability of $\widehat{\Delta}$, we next show it is asymptotically linear and identify its influence function. First define $\bar{\Delta} = \bar{\mu}_1 - \bar{\mu}_0$,

where

$$\bar{\mu}_1 = \mathbb{E}\left\{ \frac{I(T=1)Y^\dagger}{\pi(\mathbf{X};\bar{\boldsymbol{\alpha}}_1,\bar{\boldsymbol{\beta}}_1)} \right\} \text{ and } \bar{\mu}_0 = \mathbb{E}\left\{ \frac{I(T=0)Y^\dagger}{1-\pi(\mathbf{X};\bar{\boldsymbol{\alpha}}_1,\bar{\boldsymbol{\beta}}_1)} \right\},$$

with $\pi(\mathbf{x};\bar{\boldsymbol{\alpha}}_1,\bar{\boldsymbol{\beta}}_1) = \mathbb{P}(T=1 \mid \bar{\boldsymbol{\alpha}}_1^\top \mathbf{X} = \bar{\boldsymbol{\alpha}}_1^\top \mathbf{x}, \bar{\boldsymbol{\beta}}_1^\top \mathbf{X} = \bar{\boldsymbol{\beta}}_1^\top \mathbf{x})$, $\bar{\boldsymbol{\alpha}}_1$ and $\bar{\boldsymbol{\beta}}_1$ as the probability limits of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}_1$ regardless of model adequacy, and $Y^\dagger$ being defined as in (8) except that $\pi(\mathbf{x})$ is replaced by $\pi(\mathbf{x};\bar{\boldsymbol{\alpha}}_1,\bar{\boldsymbol{\beta}}_1)$. We show in Web Appendix B that, under the same requirements for α and β, the influence function for $\widehat{\Delta}$ is given by the summand of $n^{1/2}(\widehat{\Delta}_k - \bar{\Delta}_k) = \widehat{\mathcal{W}}_1 - \widehat{\mathcal{W}}_0$, where $\widehat{\mathcal{W}}_k = n^{1/2}(\widehat{\mu}_k - \bar{\mu}_k)$ for $k = 0, 1$ and

$$\widehat{\mathcal{W}}_k = n^{-1/2} \sum_{i=1}^{n} (\mathbf{v}_{\beta_1,k}^\top + \mathbf{u}_{pa,\pi,k}^\top)\boldsymbol{\varphi}_{\beta_1,i} + \mathbf{u}_{\gamma,k}^\top \boldsymbol{\varphi}_{\gamma,i} + o_p(1), \tag{14}$$

with $\mathbf{v}_{\beta_1,k} = \mathbf{0}$ when the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ is correctly specified and $\mathbf{u}_{pa,\pi,k} = \mathbf{0}$ when either the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ or imputation model $g_\xi(\boldsymbol{\gamma}^\top \mathbf{z}_\pi)$ without the utility covariate is correctly specified. Here $\boldsymbol{\varphi}_{\beta_1,i}$ and $\boldsymbol{\varphi}_{\gamma,i}$ are influence functions for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\gamma}}$ such that $n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\varphi}_{\beta_1,i} + o_p(1)$ and $n^{1/2}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\varphi}_{\gamma,i} + o_p(1)$. Accordingly, the first term in (14) represents the contribution from estimating $\boldsymbol{\beta}_1$ in the baseline outcome model $\mu_k(\mathbf{x};\boldsymbol{\beta})$ for the double-index PS appearing in the IPW weight and the utility covariate. The remaining term represents the contribution from estimating $\boldsymbol{\gamma}$ in the imputation model $g_\xi(\boldsymbol{\gamma}^\top \mathbf{z}_\pi)$. The influence function does not include terms associated with the variability in estimating $\boldsymbol{\alpha}$ in the parametric PS or for smoothing in the double-index PS, as such contributions to the expansion are of higher order when $N \gg n$ in the SS setting.

When the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ is correctly specified, the influence function (14) simplifies to

$$n^{1/2}(\widehat{\Delta} - \Delta) = n^{-1/2} \sum_{i=1}^{n} (\mathbf{u}_{\gamma,1} - \mathbf{u}_{\gamma,0})^\top \boldsymbol{\varphi}_{\gamma,i} + o_p(1),$$

where

$$\boldsymbol{\varphi}_{\gamma,i} = \left[\mathbb{E}\left\{\mathbf{Z}_{\pi,i}\mathbf{Z}_{\pi,i}\dot{g}_\xi(\bar{\boldsymbol{\gamma}}^\top \mathbf{Z}_{\pi,i})\right\}\right]^{-1} \mathbf{Z}_{\pi,i}\left\{Y_i - g_\xi(\bar{\boldsymbol{\gamma}}^\top \mathbf{Z}_{\pi,i})\right\},$$

for $\dot{g}_\xi(u) = \frac{\partial}{\partial u} g_\xi(u)|_{u=u}$. The centering of Y_i around a model approximation of $\xi_T(\mathbf{V})$ suggests $\widehat{\Delta}$ achieves efficiency gain over complete-case (CC) estimators, which neglect surrogates $\mathbf{W}$.

2.6 | Efficiency considerations

More formally, semiparametric efficiency theory establishes efficiency bounds for regular estimators of parameters of interest under specified semiparametric models (Bickel *et al*., 1998). Much of existing work involving semiparametric efficiency focus on data with iid observations. However, in our setup the observations in $\mathscr{D}$ are not exactly iid as L are constrained such that $\sum_{i=1}^N L_i = n$ for a fixed n and $\nu_n = n/N \to 0$ as $n \to \infty$. Moreover, because the distribution for the full data varies with N, as $\nu_n \to 0$, the very notion of a regular estimator is not clearly defined in this context. These issues complicate conventional applications of efficiency theory in the SS setting.

Let $\bar{\Delta}^* = \mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\}$ be strictly a functional of the observed data distribution not depending on identification assumptions (2)-(4), as in Δ. Instead of directly calculating what would be the efficient influence function for $\bar{\Delta}^*$ under the model for the full data, we take an alternative approach in which we calculate the efficient influence function for $\bar{\Delta}^*$ under an *ideal SS model* $\mathcal{M}_{SS}$ for the labeled data $\mathscr{L}$, which allows the conditional distribution of $Y \mid \mathbf{V}, T$ to be unrestricted but assumes the distribution of $(\mathbf{V}^\top, T)^\top$ is completely known. $\mathcal{M}_{SS}$ approximates the SS setting, where the size of the unlabeled data $\mathscr{U}$ is much larger than that of $\mathscr{L}$. As data in $\mathscr{L}$ itself are iid with a fixed distribution, restricting the focus to $\mathscr{L}$ avoids the complications described above. We finally calculate the asymptotic variance of $\widehat{\Delta}$ through its associated influence function and compare against the efficiency bound under $\mathcal{M}_{SS}$ to better understand the efficiency of $\widehat{\Delta}$ in context.

Following this approach, we show in Web Appendix C that the semiparametric efficiency bound for $\bar{\Delta}^*$ under $\mathcal{M}_{SS}$, with respect to a class of regular parametric submodels subject to mild regularity conditions, is $\mathbb{E}(\varphi_{\text{eff}}^2)$, where

$$\varphi_{\text{eff}} = U_\pi\{Y - \xi_T(\mathbf{V})\} \tag{15}$$

is the efficient influence function. This efficiency bound is lower than or equal to the efficiency bound in the fully nonparametric model where the distribution of $(Y, \mathbf{V}^\top, T)^\top$ is unknown. Furthermore, we show in Web Appendix C that $\widehat{\Delta}$ indeed achieves the SS efficiency bound when both the PS and imputation model, $\pi(\mathbf{x};\boldsymbol{\alpha})$ and $g_\xi(\boldsymbol{\gamma}^\top \mathbf{z}_\pi)$, are correctly specified so that $\widehat{\Delta}$ is locally semiparametric efficient. Although the distribution of $(\mathbf{V}^\top, T)^\top$ is actually not known in our data setup, the bound under the ideal model $\mathcal{M}_{SS}$ can still be achieved because $N \gg n$. The correct specification of $\mu_k(\mathbf{x};\boldsymbol{\beta})$ is not required for attaining the efficiency bound, as the bound does not involve $\mu_k(\mathbf{x})$, but its specification is still important for double-robustness in case $\pi(\mathbf{x};\boldsymbol{\alpha})$ is misspecified.

The local efficiency of $\widehat{\Delta}$ may prompt efficiency gains over nonefficient CC and SS estimators, though other estimators can also achieve this efficiency bound. For example, we show in Web Appendix C that a modification of the AIPW estimator from Davidian *et al*. (2005) also achieves the bound in our data setting, when its underlying working PS and imputation

models are correctly specified. However, its efficiency may still differ with that of $\widehat{\Delta}$ under misspecification of the working imputation model, as we consider in the simulations below. Beyond these asymptotic properties under ideal conditions, we find through the simulations that $\widehat{\Delta}$ can still achieve substantial efficiency gains over other estimators in finite samples under misspecified working models. Another issue related to misspecification is that $\mu_k(\mathbf{x};\boldsymbol{\beta})$ and $g(\boldsymbol{\gamma}^\top \mathbf{z}_\pi)$ may not be compatible with one another if nonlinear models such as logistic regression are used for either working model. Using flexible basis expansion functions in $g_\xi(\boldsymbol{\gamma}^\top \mathbf{z}_\pi)$ to more closely approximate $\xi_k(\mathbf{v})$, as suggested in (6), can mitigate such incompatibility. We find that exact compatibility of the model is not a crucial requirement for $\widehat{\Delta}$ to exhibit good performance in practice when the working models are sufficiently flexible. We next consider inference about Δ through a perturbation resampling procedure.

2.7 | Perturbation resampling

Although the asymptotic variance of $\widehat{\Delta}$ can be obtained from the influence function in (14), a direct estimate is difficult because it would require estimating complicated functionals of the data distribution. We instead propose a simple perturbation resampling procedure for inference based on Jin *et al.* (2001). Let $\mathcal{G} = \{G_i : i = 1, \ldots, N\}$ be nonnegative iid random variables with unit mean and variance that are independent of the observed data $\mathscr{D}$. We first obtained perturbed estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\alpha}}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\pi(\boldsymbol{\alpha}; \mathbf{X}_i, T_i) G_i + p^*_{\lambda_N}(\boldsymbol{\alpha}_1) \right\}$$

$$\widehat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\mu(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, T_i) G_i + p^*_{\lambda_n}(\boldsymbol{\beta}_{\{-1\}}) \right\},$$

where $p^*_{\lambda_N}(\cdot)$ and $p^*_{\lambda_n}(\cdot)$ are the corresponding penalties based weights estimated by the same perturbation procedure if data-adaptive weights are used, as in ALASSO. This leads to the perturbed DiPS estimate:

$$\widehat{\pi}^*_1(\mathbf{x}; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1) = \frac{\sum_{j=1}^N K_h(\widehat{\mathbf{S}}^*_j - \widehat{\mathbf{s}}^*) I(T_j = 1) G_j}{\sum_{j=1}^N K_h(\widehat{\mathbf{S}}^*_j - \widehat{\mathbf{s}}^*) G_j},$$

where $\widehat{\mathbf{S}}_j = (\widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)^\top \mathbf{X}_j$ and $\widehat{\mathbf{s}}^* = (\widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)^\top \mathbf{x}$ are the perturbed bivariate scores. We then obtain the perturbed estimator $\widehat{\boldsymbol{\gamma}}^*$ as the solution to

$$n^{-1} \sum_{i=1}^{n} \mathbf{Z}_{\widehat{\pi}^*,i} \left\{ Y_i - g_\xi(\boldsymbol{\gamma}^\top \mathbf{Z}_{\widehat{\pi}^*,i}) \right\} G_i + \lambda_n \boldsymbol{\gamma}_{\{-1\}} = \mathbf{0},$$

where $\widehat{\pi}^*$ specifies that the imputations use utility covariates that plug in $\widehat{\pi}^*(\mathbf{x}; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)$. Finally, we calculate the perturbed $\mathrm{SS}_{\mathrm{DR}}$ estimator as $\widehat{\Delta}^* = \widehat{\mu}^*_1 - \widehat{\mu}^*_0$, where

$$\widehat{\mu}^*_1 = \left\{ \sum_{i=1}^{N} \frac{I(T_i = 1) G_i}{\widehat{\pi}(\mathbf{X}_i; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)} \right\}^{-1} \left\{ \sum_{i=1}^{N} \frac{I(T_i = 1) \widehat{Y}^*_i G_i}{\widehat{\pi}(\mathbf{X}_i; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)} \right\}$$

$$\text{and } \widehat{\mu}^*_0 = \left\{ \sum_{i=1}^{N} \frac{I(T_i = 0) G_i}{1 - \widehat{\pi}(\mathbf{X}_i; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)} \right\}^{-1} \times \left\{ \sum_{i=0}^{N} \frac{I(T_i = 0) \widehat{Y}^*_i G_i}{1 - \widehat{\pi}(\mathbf{X}_i; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)} \right\},$$

with $\widehat{Y}^*_i = g_\xi(\widehat{\boldsymbol{\gamma}}^{*\top} \mathbf{Z}_{\widehat{\pi}^*,i})$. It can be shown using arguments from Jin *et al.* (2001) that the asymptotic distribution of $n^{1/2}(\widehat{\Delta} - \bar{\Delta})$ coincides with that of $n^{1/2}(\widehat{\Delta}^* - \widehat{\Delta}) \mid \mathscr{D}$. In this scheme, estimation of the initial PS model coefficients $\widehat{\boldsymbol{\alpha}}^*$, the DiPS $\widehat{\pi}^*(\mathbf{x}; \widehat{\boldsymbol{\alpha}}^*_1, \widehat{\boldsymbol{\beta}}^*_1)$, and the final IPW estimators $\widehat{\mu}^*_1$ and $\widehat{\mu}^*_0$ does not technically need to be perturbed as they are estimated based on data from $\mathscr{U}$. Consequently, their contributions to the asymptotic variance is of higher order when $N \gg n$. However, we found that not perturbing these steps can have some impact on the standard error estimation in finite samples if N is not yet very large relative to n and perturbed these steps by default. We approximate the standard error of $\widehat{\Delta}$ based on the empirical standard deviation, or, as a robust alternative, the mean absolute deviation (MAD), of a large number of samples of $\widehat{\Delta}^*$ and construct confidence intervals (CI) based on the empirical percentiles.

3 | SIMULATIONS

We assessed through simulations the finite samples bias, standard error (SE), root mean square error (RMSE), and relative efficiency (RE) of our proposed estimator ($\mathrm{SS}_{\mathrm{DR}}$) compared to alternative estimators. In separate simulations we also examined the performance of the perturbation procedure for inference based on $\mathrm{SS}_{\mathrm{DR}}$. For $\mathrm{SS}_{\mathrm{DR}}$, we specified $\mathbf{h}(\cdot)$ in the imputation model in (6) as natural cubic splines with six knots specified at uniform quantiles. Ridge regression with the tuning parameter chosen by cross-validation on the deviance was used for regularization in (7). Although it is unclear whether cross-validation for choosing the tuning parameter satisfies with high probability the $\lambda_n = o(n^{-1/2})$ condition, we found that it leads to good performance in finite samples and that the simulation results were not sensitive to the choice of the tuning parameter (Web Appendix D). ALASSO with initial weights estimated by ridge regression and tuning parameter chosen by minimizing a modified BIC criteria (Minnier *et al.*, 2011) was used for estimating $\widehat{\boldsymbol{\alpha}}$

and $\widehat{\boldsymbol{\beta}}$. A plug-in estimate was used for the bandwidth in the smoothing for the double-index PS (Cheng *et al*., 2019). Prior to smoothing, the components of $\widehat{\mathbf{S}}$ were standardized and transformed by a probability integral transform based on the normal cumulative distribution function to induce approximately uniformly distributed inputs, which can improve finite-sample performance (Wand *et al*., 1991). As we focused on binary outcomes, we specified logistic link functions $g_{\xi}(u) = g_{\pi}(u) = g_{\mu}(u) = 1/(1+e^{-u})$ for the working models in (6) and (10).

For comparison, we considered the standard complete-case AIPW estimator (Lunceford and Davidian, 2004) based on $\mathscr{L}$ only (CC_{AIPW}). We also considered the estimator from Davidian *et al*. (2005), which we adapted to our setting by replacing instances of the treatment randomization probability with PS estimates $\pi(\mathbf{X}_i;\widehat{\boldsymbol{\alpha}})$ (SS_{AIPW}). SS_{AIPW} also leverages the surrogate data in $\mathscr{U}$ alongside labeled data $\mathscr{L}$ to facilitate estimation of the ATE and is locally efficient, as discussed in Section 2.6. It is also DR in that it is consistent for Δ if either the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ or outcome regression model $\mu_k(\mathbf{x};\boldsymbol{\beta})$ is correctly specified.

To mimic the EHR data, we considered the case with Y as binary and $\mathbf{W}$ as count variables. In all scenarios, data were generated according to $\mathbf{X} \sim N\{\mathbf{0}, \sigma_x^2(1-\rho_x)\mathbf{I} + \sigma_x^2\rho_x\}$, $T \mid \mathbf{X} \sim \text{Bernoulli}\{\pi(\mathbf{X})\}$, $Y \mid \mathbf{X}, T \sim \text{Bernoulli}\{\mu_T(\mathbf{X})\}$, and $\mathbf{W} = \lfloor \boldsymbol{\Gamma}(1, \mathbf{X}^{\top}, T, Y)^{\top} + \boldsymbol{\epsilon} \rfloor$, where $\boldsymbol{\epsilon} \sim N\{\mathbf{0}, \sigma_w^2(1-\rho_w)\mathbf{I} + \sigma_w^2\rho_w\}$ and $\lfloor\cdot\rfloor$ is the floor function. Initially we considered simulating data that roughly resembled the EHR data example in terms of model parameters but were simple enough to be more broadly relevant. To this end, we considered $p_x = 10$ baseline covariates and $p_w = 5$ surrogates, with variances and correlations $\sigma_x^2 = 1$, $\rho_x = .2$, $\sigma_w^2 = 5$, $\rho_w = .2$, and $\boldsymbol{\Gamma}_{5\times 13} = (\mathbf{0}_{5\times 1}, .1\mathbf{1}_{5\times 5}, -.1\mathbf{1}_{5\times 5}, .1\mathbf{1}_{5\times 1}, \mathbf{g})$. These simulations were varied over different model specifications, predictive strength of surrogates $\mathbf{W}$, and sample sizes. The imputation model $\xi(\mathbf{v};\boldsymbol{\gamma})$ was misspecified throughout, and either the PS model $\pi(\mathbf{x};\boldsymbol{\alpha})$ or baseline outcome model $\mu_k(\mathbf{x};\boldsymbol{\beta})$ were potentially misspecified by setting the true $\pi(\mathbf{x})$ and $\mu_k(\mathbf{x})$ to be double-index models that allow for pairwise interactions:

(1) *Both correct:* $\mu_k(\mathbf{x}) = g_{\mu}(\beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{x} + \beta_2 k)$,

$$\pi(\mathbf{x}) = g_{\pi}(\alpha_0 + \boldsymbol{\alpha}_1^{\top}\mathbf{x})$$

(2) *Misspecifiedμ:* $\mu_k(\mathbf{x}) = g_{\mu}\left\{\beta_0 + \boldsymbol{\beta}_{1[1]}^{\top}\mathbf{x}(\boldsymbol{\beta}_{1[2]}^{\top}\mathbf{x} + 1) + \beta_2 k\right\}$,

$$\pi(\mathbf{x}) = g_{\pi}(\alpha_0 + \boldsymbol{\alpha}_1^{\top}\mathbf{x})$$

(3) *Misspecifiedπ:* $\mu_k(\mathbf{x}) = g_{\mu}(\beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{x} + \beta_2 k)$,

$$\pi(\mathbf{x}) = g_{\pi}\left\{\alpha_0 + \boldsymbol{\alpha}_{1[1]}^{\top}\mathbf{x}(\boldsymbol{\alpha}_{1[2]}^{\top}\mathbf{x} + 1)\right\},$$

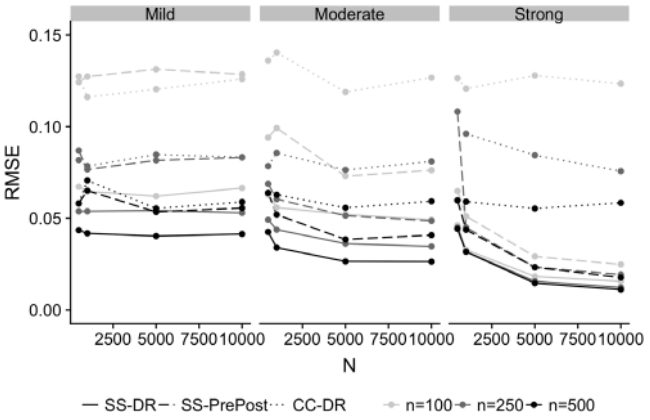

FIGURE 1 Root mean square error (RMSE) of CC_{AIPW}, SS_{AIPW}, and SS_{DR} by strength of the surrogates $\mathbf{W}$ (Mild, Moderate, Strong) over different sample sizes

where $g_{\mu}(u) = g_{\pi}(u) = 1/(1+e^{-u})$ and parameter values were

$$\alpha_0 = -.3, \quad \boldsymbol{\alpha}_1 = .35\mathbf{1}_{1\times 10}, \quad \beta_0 = -.65,$$
$$\boldsymbol{\beta}_1 = (1\mathbf{1}_{1\times 3}, .5\mathbf{1}_{1\times 3}, -1.15, -1\mathbf{1}_{1\times 3})^{\top}$$
$$\boldsymbol{\alpha}_{1[1]} = .5(0, .35, 0, .35, 0, .35, 0, .35, 0, .35)^{\top},$$
$$\boldsymbol{\alpha}_{1[2]} = (.35, 0, .35, 0, .35, 0, .35, 0, .35, 0)^{\top}$$
$$\boldsymbol{\beta}_{1[1]} = .5(1, 0, 1, 0, .5, 0, -.5, 0, -1, 0)^{\top},$$
$$\boldsymbol{\beta}_{1[2]} = (0, .5, 0, .5, 0, .5, 0, .5, 0, .5)^{\top}.$$

We considered $(2.1, 2.1, 1, 0, 0)^{\top}$, $(5, 5, 2.5, 0, 0)^{\top}$, and $(15, 10, 5, 2.5, 0)^{\top}$ for $\mathbf{g}$ to model cases where $\mathbf{W}$ has mild, moderate, and strong predictive strength, respectively. The outcome prevalence was approximately 45% in all scenarios, and the treatment prevalence was also 45%, except in the misspecified π scenario where it was approximately 53%. Sample sizes were varied over $n = 100, 250, 500$ and $N = 500, 1000, 5000, 10\,000$ in a factorial design to show separate effects of increasing n and N. To more closely approximate the data in the EHR data application, in a separate simulation, we generated data under the "both correct" setting using coefficient estimates from fitting corresponding working models in the EHR data as the values set for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}$. For these simulations, we only considered the size $n = 100$, $N = 1000$. The results in each scenario are summarized from 1000 simulated datasets.

Table 1 presents the bias, SE, and RMSE across misspecification scenarios with $\mathbf{W}$ at moderate strength. SS_{DR}, SS_{AIPW}, and CC_{AIPW} exhibits low bias that diminishes to zero as sample size increased under all three scenarios, verifying their double-robustness. Decreases in bias for SS_{DR} appears to be largely driven by increasing n. Figure 1 presents the RE under the correctly specified scenario varying the strength of

TABLE 1 Bias, SE, and RMSE of estimators under different model misspecification scenarios over 1000 simulated datasets

			Both Correct			Misspecified μ			Misspecified π		
n	N	Estimator	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
100	1000	CC_{AIPW}	−0.001	0.140	0.140	−0.003	0.176	0.176	0.002	0.088	0.088
100	1000	SS_{AIPW}	−0.006	0.099	0.099	−0.009	0.110	0.110	−0.003	0.058	0.058
100	1000	SS_{DR}	−0.007	0.055	0.056	−0.016	0.061	0.063	−0.008	0.046	0.047
100	10 000	CC_{AIPW}	0.005	0.127	0.127	0.008	0.172	0.172	0.001	0.083	0.083
100	10 000	SS_{AIPW}	−0.003	0.076	0.076	−0.003	0.113	0.113	−0.005	0.053	0.053
100	10 000	SS_{DR}	−0.005	0.049	0.049	−0.014	0.056	0.058	−0.009	0.040	0.042
500	1000	CC_{AIPW}	−0.002	0.063	0.063	−0.000	0.083	0.083	0.001	0.034	0.034
500	1000	SS_{AIPW}	−0.003	0.052	0.052	−0.001	0.072	0.072	−0.000	0.029	0.029
500	1000	SS_{DR}	−0.003	0.034	0.034	−0.002	0.044	0.044	−0.003	0.027	0.028
500	10 000	CC_{AIPW}	−0.002	0.059	0.059	0.002	0.081	0.081	−0.000	0.035	0.035
500	10 000	SS_{AIPW}	−0.002	0.041	0.041	−0.001	0.048	0.048	−0.001	0.023	0.023
500	10 000	SS_{DR}	−0.003	0.026	0.026	−0.002	0.033	0.033	−0.004	0.020	0.020

association between $\mathbf{W}$ and Y. Generally, SS_{DR} uniformly achieves the lowest RMSE for a given n. Increasing N while fixing n improves the RMSE for SS_{DR} and SS_{AIPW}, but the improvements are limited by n, which drives the asymptotic variance in the SS regime as shown in the asymptotic analysis. The benefit of additional unlabeled data varies with the strength of $\mathbf{W}$. The RMSE for SS_{DR} does not improve much with larger N for a fixed n when $\mathbf{W}$ is weakly correlated with Y but improves greatly when $\mathbf{W}$ are strongly correlated.

Figure 2 presents the RE of various estimators relative to SS_{DR} across misspecification scenarios with moderate $\mathbf{W}$. SS_{DR} is more efficient than both CC_{AIPW} and SS_{AIPW} across misspecification settings and sample sizes. It gains over CC_{AIPW} as it makes use of the unlabeled data $\mathcal{U}$. The gains over SS_{AIPW} suggest that SS_{DR} can be more efficient under misspecification of the working imputation model. In other simulations not presented we found that SS_{DR} has similar efficiency with SS_{AIPW} under a correctly specified imputation model, as expected since both are locally efficient. SS_{DR} may also achieve efficiency gains relative to other estimators involving PS weighting in finite samples when the true PS are more extreme. The calibrated estimate $\widehat{\pi}(\mathbf{x}; \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1)$ used in SS_{DR} pulls estimates of the PS away from 0 or 1, which can lead to more stable final estimates when $\widehat{\pi}(\mathbf{x}; \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1)$ is used in reweighting. Finally, SS_{DR} may exhibit some efficiency gains over SS_{AIPW} in finite samples from using regularization for estimating the nuisance parameters, whereas SS_{AIPW} uses unregularized maximum likelihood estimators in our implementation. In the data application scenario, SS_{DR} was significantly more efficient, having an RMSE of .03 compared to .15 and .12 for CC_{AIPW} and SS_{AIPW}, suggesting the strength of surrogates are strong in the EHR data.

To implement the perturbation procedure, we used the weights $G_i \sim 4 \times \text{Beta}(.5, 1.5)$ and 1000 sets of $\mathcal{G}$ for SE and CI estimation. We considered evaluating the perturbations only in the scenario where both $\mu_k(\mathbf{x}; \boldsymbol{\beta})$ and $\pi(\mathbf{x}; \boldsymbol{\alpha})$ were correctly specified and $\mathbf{W}$ had moderate predictive strength. The results are presented in Table 2. In both small and large samples, the SEs estimated by the standard deviation and by MAD approximated well the empirical standard error. The coverage of the percentiles were also close to nominal levels, albeit slightly conservative. Results from four of the simulation iterations for when $n = 100$, $N = 1000$ and from eight of the iterations when $n = 250$, $N = 5000$ were omitted as the simulations timed out from prolonged computational time.

4 | EHR DATA APPLICATION

We applied SS_{DR} and the alternative estimators to compare the rates of treatment response to two biologic agents for treating inflammatory bowel disease (IBD) using the EMRs from Partners' Healthcare. Though the efficacy and effectiveness of adalimumab (ADA) and infliximab (IFX) for the management of IBD have been established individually, few studies have offered a direct comparison. Consequently, the choice of treatment in practice is often influenced by factors other than comparative performance (Ananthakrishnan *et al.*, 2016). Randomized trials may be unfeasible due to the large number of patients that would be needed to detect the presumed small treatment difference, and other observational data lack detailed clinical information needed to ascertain meaningful outcomes. EHRs are thus uniquely positioned to provide evidence on the comparative effectiveness of these two therapies.

The data we considered consisted of $N = 1243$ total IBD patients, including 200 who initiated treatment with ADA and 1043 with IFX. Through chart review by a gastroenterologist, a random subset of $n = 117$ records were labeled with the true treatment response status (responder vs nonresponder)

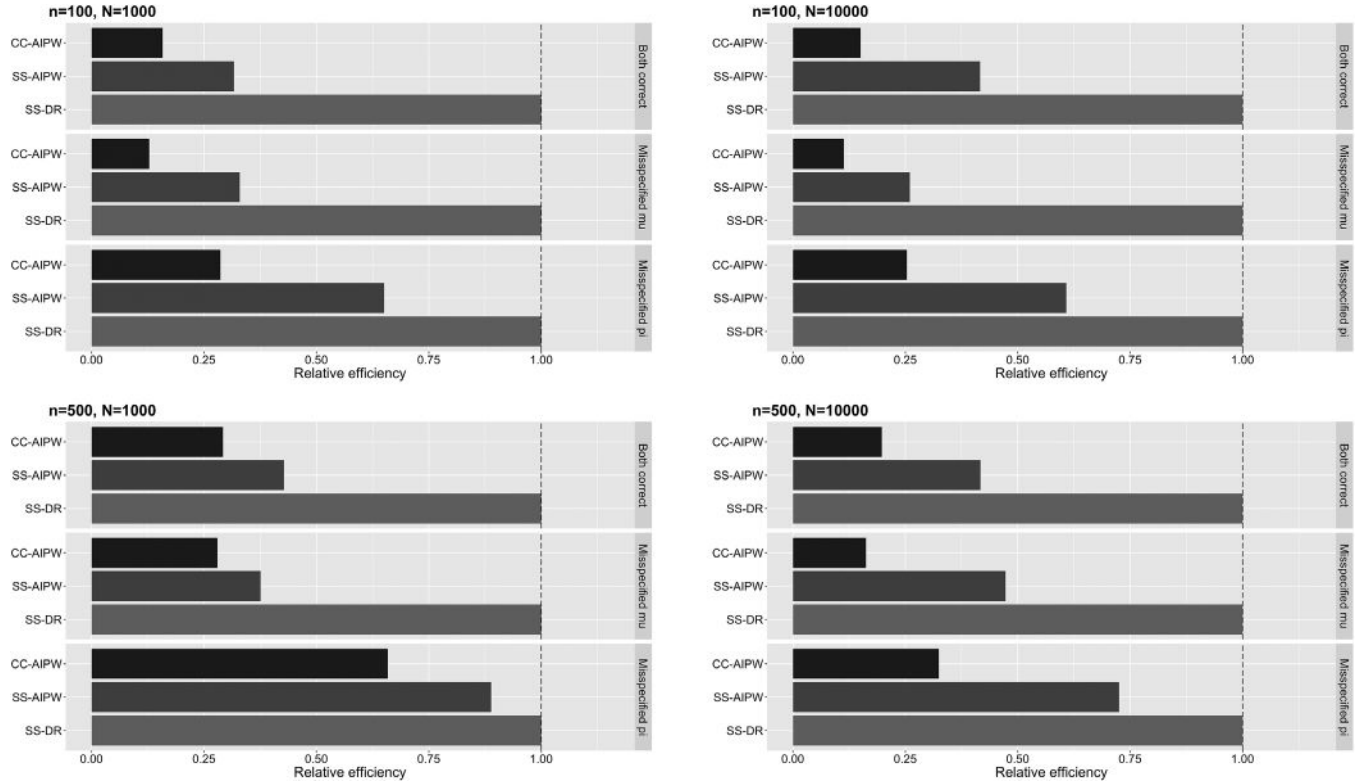

FIGURE 2 RE of estimators, defined as the ratio of mean square errors (MSE) relative to SS_{DR}, by model misspecification scenarios for different sample sizes over 1000 simulated datasets. Higher values of RE denotes greater efficiency (lower MSE) relative to SS_{DR}

TABLE 2 Performance of perturbation resampling for SS_{DR} in 1000 simulated datasets when both $\mu_k(\mathbf{x};\boldsymbol{\beta})$ and $\pi(\mathbf{x};\boldsymbol{\alpha})$ are correctly specified

Size	Bias	Emp SE	ASE	ASE_{MAD}	RMSE	Coverage
$n = 100$, $N = 1000$	−0.002	0.055	0.052	0.052	0.055	0.963
$n = 250$, $N = 5000$	−0.002	0.034	0.034	0.034	0.034	0.963

Abbreviations: ASE: average of estimated SE based on the standard deviation of perturbed estimates; ASE_{MAD}: average of SE based on MAD of perturbed estimates; Coverage: coverage of 95% percentile CIs; RMSE: root-mean square error; SE: empirical SE of SS_{DR} over simulated datasets.

within 1 year of treatment initiation. We included 12 baseline covariates to adjust for confounding in **X**, including demographics, comorbidities, prior utilization, and inflammation biomarker levels. We also selected 35 post-treatment surrogates for **W**, comprising counts of NLP mentions of clinically relevant terms (eg, "bleeding," "fistula," and "tenesmus") within 1 year of initiation. The transformation $u \mapsto \log(1+u)$ was applied to all count variables in **V** to mitigate instability in the estimation due to skewness in their distributions. Nonparametric bootstrap was used to estimate SEs and CIs for the alternative estimators and perturbation for SS_{DR}, using the MAD of resampled estimates as an robust estimator of the SEs. In addition we calculated two-sided *P*-values based on inverting percentile CIs from the resampled estimates, using the equivalence between significance tests and confidence sets (Liu and Singh, 1997).

As shown in Table 3, the point estimates generally indicate that patients receiving ADA experienced lower rates of treatment response, after adjustment for confounding. SS_{DR} is estimated to achieve more than 600% efficiency gain over CC estimators and 450% efficiency gain over the other SS estimators based on the estimated variances. It is the only estimator that exhibits a difference that is significant at the .05 level,

TABLE 3 Point and SE estimates based on MAD for the ATE of ADA versus IFX, with respect to 1-year treatment response rate, among IBD patients in EMR data based on various methods, including the naive CC estimator (CC_{Naive}) that completely ignores confounding bias

Estimator	Estimate	SE	95% CI (Pct)	*P*-value
CC_{Naive}	0.014	0.099	(−0.201, 0.177)	0.822
CC_{AIPW}	−0.125	0.153	(−0.416, 0.164)	0.592
SS_{AIPW}	0.033	0.109	(−0.265, 0.180)	0.778
SS_{DR}	−0.067	0.036	(−0.164, −0.002)	0.044

Note. 95% CIs are percentile-based CIs from resampling and *P*-values are for testing $H_0 : \Delta = 0$ based on inverting percentile CIs.

suggesting that patients receiving IFX experience a slightly higher rate of response to treatment.

5 | DISCUSSION

This paper developed a robust and efficient estimator for the ATE in an SS setting where the true outcome is labeled for a vanishingly small proportion of the entire set of observations. The estimator adopts an imputation approach to leverage surrogate data from $\mathcal{U}$ to improve efficiency that is robust to misspecification of the imputation model. It is DR, locally semiparametric efficient under an ideal SS semiparametric model, and demonstrated to be more efficient than CC and other estimators that leverage $\mathcal{U}$ in finite samples.

We have assumed that the true outcomes Y are labeled completely at random, which may be reasonable if investigators control the labeling. But this assumption could be restrictive if labeling was stratified by some known factors or if some records that are available were not labeled for research purposes. One possible approach to address the case where Y are missing at random is to apply weighting or semiparametric efficient methods (Robins *et al*., 1994) to the estimating equation when estimating $\boldsymbol{\gamma}$ in (7). Other refinements to our proposed approach are possible. For example, in the case where $\mathbf{W}$ is high dimensional, the group LASSO (Yuan and Lin, 2006), where the basis expansion functions for each surrogate variable are grouped together, can also potentially be used to improve efficiency in finite-samples. It would also be of interest to extend the theoretical results to the case where p_x and p_w are allowed to diverge with n.

ACKNOWLEDGMENTS

The authors thank Ray Liu, Eric Tchetgen Tchetgen, Rajarshi Mukherjee, and James Robins for helpful discussions as well as the editor, associate editor, and two referees for their insightful feedback and suggestions. Much of this work was done when the first author was a graduate student at Harvard University. This work was supported by National Institutes of Health grants T32CA009337, R21CA242940, and R01HL089778. The views expressed in this article are those of the authors and do not necessarily reflect the views of the Department of Veterans Affairs.

ORCID

David Cheng https://orcid.org/0000-0002-2816-6585
Ashwin N. Ananthakrishnan
https://orcid.org/0000-0002-9436-1821

REFERENCES

Ananthakrishnan, A.N., Cagan, A., Cai, T., Gainer, V.S., Shaw, S.Y., Savova, G., Churchill, S., Karlson, E.W., Kohane, I., Liao, K.P. and Murphy, S.N., (2016) Comparative effectiveness of infliximab and adalimumab in Crohn's disease and ulcerative colitis. *Inflammatory Bowel Diseases*, 22, 880–885.

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

Benkeser, D., Carone, M., van der Laan, M.J. and Gilbert, P.B. (2017) Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104, 863–880.

Bickel, P.J., Klaassen, C.A., Ritov, Y., Klaassen, J. and Wellner, J.A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York, NY: Springer.

Chen, S.X., Leung, D.H.Y. and Qin, J. (2003) Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association*, 98, 1052–1062.

Cheng, D., Chakrabortty, A., Ananthakrishnan, A.N. and Cai, T. (2019) Estimating average treatment effects with a response-informed calibrated propensity score. *Biometrics*, 76, 767–777.

Davidian, M., Tsiatis, A.A. and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20, 261.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96, 1348–1360.

Jin, Z., Ying, Z. and Wei, L.-J. (2001) A simple resampling method by perturbing the minimand. *Biometrika*, 88, 381–390.

Liu, R.Y. and Singh, K. (1997) Notions of limiting p values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266–277.

Lunceford, J.K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23, 2937–2960.

Minnier, J., Tian, L. and Cai, T. (2011) A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106, 1371–1382.

Pepe, M.S. (1992) Inference using surrogate outcome data and a validation sample. *Biometrika*, 79, 355–365.

Robins, J.M. (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7, 1393–1512.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J.M. (2012) Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 439–456.

Rotnitzky, A., Robins, J.M. and Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93, 1321–1339.

van der Laan, M.J. and Rubin, D. (2006) Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2, 11.

Wand, M.P., Marron, J.S. and Ruppert, D. (1991) Transformations in density estimation. *Journal of the American Statistical Association*, 86, 343–353.

Williamson, E., Forbes, A. and Wolfe, R. (2012) Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in Medicine*, 31, 4382–4400.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, Z., Liu, W., Zhang, B., Tang, L. and Zhang, J. (2016) Causal inference with missing exposure information: Methods and applications to an obstetric study. *Statistical Methods in Medical Research*, 25, 2053–2066.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 2 and 3 are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Cheng D, Ananthakrishnan AN, Cai T. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*. 2021;77:413–423. https://doi.org/10.1111/biom.13298