Efficient and Robust Semi-Supervised Estimation of Average Treatment Effects in Electronic Medical Records Data

David Cheng¹, Ashwin Ananthakrishnan², and Tianxi Cai¹,*

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A.
²Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, U.S.A.

*email: tcai@hsph.harvard.edu

Summary: There is strong interest in conducting comparative effectiveness research (CER) in electronic medical records (EMR) to evaluate treatment strategies among real-world patients. Inferring causal effects in EMR data, however, is challenging due to the lack of direct observation on pre-specified gold-standard outcomes, in addition to the observational nature of the data. Extracting gold-standard outcomes often requires labor-intensive medical chart review, which is unfeasible for large studies. While one may impute outcomes and estimate average treatment effects (ATE) based on imputed data, naive imputations may lead to biased ATE estimators. In this paper, we frame the problem of estimating the ATE in a semi-supervised learning setting, where a small set of observations is labeled with the true outcome via manual chart review and a large set of unlabeled observations with features predictive of the outcome are available. We develop an imputation-based approach for estimating the ATE that is robust to misspecification of the imputation model. This allows information from the predictive features to be safely leveraged to improve the 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. We show that it is locally semiparametric efficient under an ideal semi-supervised model where the distribution of 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 to compare rates of treatment response to two biologic agents for treating inflammatory bowel disease using EMR data from Partner’s Healthcare.

Key words: Causal inference; Double-robustness; Missing data; Propensity score; Robust imputation, Semiparametric efficiency; Surrogate outcomes.
1. Introduction

Electronic medical records (EMR) aggregate rich patient-level data that are routinely collected during patient care. Since they include large samples in broad populations, EMRs have become a valuable data source for conducting comparative effectiveness research (CER) and identifying optimal treatment strategies among real-world patients (Fiks et al., 2012; Manion et al., 2012). However, many challenges arise when performing CER using EMR data (Hersh et al., 2013). Beyond the usual issue of confounding in observational data, a primary challenge is the lack of direct observation on a true clinical outcome of interest \( Y \). In contrast with clinical trials or traditional observational studies, EMR data are not collected to evaluate any specific pre-specified outcome. This is frequently ignored, whereby researchers often explicitly or implicitly rely on “surrogate variables” \( W \) abstracted from codified (e.g. billing codes) or narrative (e.g. physician notes) data to approximate \( Y \) with some imputed outcome \( Y^\dagger = g(W) \). But the accuracy of \( Y^\dagger \) in approximating \( Y \) is usually unclear, and reliance on naive imputations \( Y^\dagger \) can lead to biased estimates of the average treatment effect (ATE) on the actual outcome of interest \( Y \).

To address this issue, EMR studies often undertake a manual chart review process by domain experts to label records with a “gold-standard” \( Y \). But labeling is a costly and time-consuming process, which is effectively unfeasible when scaling a study up to a large sample. An alternative approach used in practice is to label a relatively small number of records and use the labeled data \( \mathcal{L} \) to build an imputation model for \( Y \) based on features extracted automatically from the EMR, including \( W \) (Ananthakrishnan et al., 2016). This approach is attractive because it uses \( Y \) to build the imputation model and avoids comprehensive labeling of all records. But it remains unclear when a resulting estimator of the ATE will be valid or efficient, especially under possible misspecification of the imputation model.

To the best of our knowledge, there have been no methods developed for this problem in the
context of EMR data. This problem can be cast in terms of surrogate outcomes data, where \( \mathbf{W} \) can be regarded as surrogates for \( Y \) and \( \mathcal{L} \) can be regarded as a validation sample. A variety of methods have been developed in this data setting for estimating regression parameters (Pepe, 1992; Reilly and Pepe, 1995; Chen, 2000; Chen and Chen, 2000) and solutions to estimating equations (Chen et al., 2003, 2008). These methods tend to assume a univariate surrogate with low dimensional baseline covariates \( \mathbf{X} \) and have not been adapted to estimate causal effects in observational data. More generally, this problem can be regarded as a missing data problem in which \( Y \) is missing in the large set of unlabeled data \( \mathcal{U} \), for which semiparametric efficiency theory can potentially be applied (Robins et al., 1994, 1995; Robins and Rotnitzky, 1995). This approach has been used to develop robust and efficient estimators of the ATE in clinical trial settings with availability of surrogates (Davidian et al., 2005) and observational data with missingness in variables besides \( Y \) (Williamson et al., 2012; Zhang et al., 2016). But these methods assume the proportion of missingness is bounded away from 0, whereas in the EMR setting, \( \mathcal{L} \) is always so much smaller than \( \mathcal{U} \) that the proportion of missingness should be regarded as tending to 0 asymptotically. This in turn changes the semiparametric efficiency considerations. Moreover, these methods involve parametric modeling of propensity score (PS) and outcome models, and typical implementations based on logistic and linear regression models have poor performance if they are misspecified or if the number of covariates is not small. Other methods for handling missing data in non-EMR settings based on imputation models (Little and An, 2004; An and Little, 2008) are related to our proposed approach but do not address estimation of causal treatment effects.

In this paper, we propose a semi-supervised (SS) estimator for the ATE (SSDR) based on an imputation followed by inverse probability weighting (IPW) that is doubly-robust and semiparametric efficient. The imputations are constructed such that the resulting estimator is robust to misspecification of the imputation model, enabling \( \mathcal{U} \) to be safely used in
improving estimation. We further employ a double-index propensity score \cite{Cheng et al., 2017} for additional robustness and efficiency gain. The remainder of the paper is organized as follows. We formalize the SS estimation problem in Sections 2.1-2.2 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 Partner’s 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 that could be modeled by a generalized linear model (GLM), such as a binary, ordinal, or continuous response, $T \in \{0,1\}$ a binary treatment, $X$ a $p_x$-dimensional vector of pre-treatment baseline covariates, $W$ a $p_w$-dimensional vector of post-treatment surrogate variables that are potentially predictive of $Y$, and $V = (W^T, X^T)^T$. The labeled data consists of $n$ independent and identically distributed (iid) observations $\mathcal{L} = \{(Y_i, T_i, V_i^T)^T : i = 1, \ldots, n\}$, while the unlabeled data consists of $N$ iid observations without $Y$, $\mathcal{U} = \{(T_i, V_i^T)^T : i = n+1, \ldots, N\}$, with $\mathcal{U} \perp \perp \mathcal{L}$. In the SS setting $N \gg n$ so that $\nu_n = n/N \to 0$ as $n \to \infty$. We assume that the labeled observations were randomly selected so that $Y$ is missing completely at random (MCAR) from observations in $\mathcal{U}$.

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 $\mathcal{Z} = \mathcal{L} \cup \mathcal{U}$ we want to estimate the ATE:

\[
\Delta = \mathbb{E}\{Y^{(1)}\} - \mathbb{E}\{Y^{(0)}\} = \mu_1 - \mu_0.
\]
We require the following standard assumptions to identify $\Delta$:

\[ Y = TY^{(1)} + (1 - T)Y^{(0)} \]  

\[ (Y^{(1)}, Y^{(0)}) \perp T \mid X \]  

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

where $\pi(x) = P(T = 1 \mid X = x)$ is the PS and $f(x)$ 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):

\[ \Delta = E\{\mu_1(X) - \mu_0(X)\} = E \left\{ \frac{I(T = 1)Y}{\pi(X)} - \frac{I(T = 0)Y}{1 - \pi(X)} \right\}, \]

where $\mu_k(x) = E(Y \mid X = x, T = k)$ 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 scarce but surrogates $W$ are available, a further decomposition can potentially be helpful:

\[ \Delta = E[E\{\xi_1(V) \mid X, T = 1\} - E\{\xi_0(V) \mid X, T = 0\}] = E \left\{ \frac{I(T = 1)\xi_1(V)}{\pi(X)} - \frac{I(T = 0)\xi_0(V)}{1 - \pi(X)} \right\}, \]

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

Let $U_\pi = I(T = 1)/\pi(X) - I(T = 0)/\{1 - \pi(X)\}$ denote a utility covariate given $\pi(x)$, assumed momentarily to be known. Suppose we postulate a parametric working model, possibly misspecified, for $\xi_k(v)$:

$$
\xi_T(V) = g_\xi(\gamma_0 + \gamma_1^T h(V) + \gamma_2 T + \gamma_3 U_\pi) = g_\xi(\gamma^T Z_\pi),
$$

where $\gamma = (\gamma_0, \gamma_1^T, \gamma_2, \gamma_3)^T$, $Z_\pi = (1, V^T, T, U_\pi)$, $g_\xi(\cdot)$ is a specified link function, and $h(\cdot)$ is a vector of fixed basis expansion functions that can incorporate nonlinear effects. We estimate $\gamma$ as $\hat{\gamma}$, the solution to a penalized estimating equation with ridge regularization:

$$
n^{-1} \sum_{i=1}^n Z_{\pi,i} \{Y_i - g_\xi(\gamma^T Z_{\pi,i})\} + \lambda_n \gamma_{\{\gamma\}} = 0,
$$

where $\gamma_{\{\gamma\}}$ denotes the vector $\gamma$ excluding the first element $\gamma_0$ and $\lambda_n = o(n^{-1/2})$ is a tuning parameter chosen such that $\hat{\gamma}$ has $n^{-1/2}$ convergence rate. In particular, this class of estimators includes ridge estimators for GLMs based on exponential families with canonical link functions. Other regularization penalties besides the ridge penalty can also be used, as long as $\hat{\gamma}$ maintains a $n^{-1/2}$ convergence rate. Using the fact that $Y$ are MCAR, standard arguments can be used to show that $\hat{\gamma} \xrightarrow{P} \gamma$ where $\gamma$ solves:

$$
\mathbb{E} \{Z_\pi \{Y - g_\xi(\gamma^T Z_\pi)\}\} = 0,
$$

with the expectation being taken over the entire population and not restricted only to the labeled population. Specifically, for $Y^\dagger = g_\xi(\gamma^T Z_\pi)$, since $Z_\pi$ includes $U_\pi$ this implies that:

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

This suggests that a standard IPW estimator based on the imputed outcome $Y^\dagger$ has the same limit asymptotically 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 a consistent estimator of $Y^\dagger$.

In practice $\pi(x)$ also needs to be estimated, which is typically done through parametric
modeling such as logistic regression. When \( \pi(x) \) is estimated by an estimator \( \hat{\pi}(x) \), the IPW estimator discussed above will be consistent for \( \Delta \) if \( \hat{\pi}(x) \) is consistent for \( \pi(x) \) but otherwise could be biased if the parametric model for \( \pi(x) \) is misspecified. Alternatively, similar arguments can be used to construct imputations \( Y^\dagger \) that could be substituted for \( Y \) in an outcome regression estimator and still be robust to misspecification of the imputation model. However, such an approach would then require correct specification of an outcome regression model given baseline covariates for \( \mu_k(x) \) to be consistent for \( \Delta \). In the following we propose an IPW approach but weighting with a double-index PS (DiPS) \citep{Cheng2017}. The resulting IPW estimator will be doubly-robust in that it will be consistent for \( \Delta \) when a model for either \( \pi(x) \) or \( \mu_k(x) \) is correctly specified. Whereas \citep{Cheng2017} considered only the scenario where \( Y \) is fully observed, the present paper uses the double-index PS to develop a novel ATE estimator in the SS setting. Using the double-index PS is not essential in that an augmented IPW estimator \citep{Robins1994,Robins1995} can potentially be used to achieve double-robustness as well. We use the double-index PS for the clarity of the construction, which directly follows the line of reasoning of the robust imputations.

### 2.4 Doubly-Robust IPW Based on the Double-Index PS

Suppose we postulate the following working parametric models for \( \pi(x) \) and \( \mu_k(x) \):

\[
\pi(X) = g_\pi(\alpha_0 + \alpha_1^T X) = \pi(X; \alpha) \tag{9}
\]

\[
\mu_T(X) = g_\mu(\beta_0 + \beta_1^T X + \beta_2 T) = \mu_T(X; \beta), \tag{10}
\]

where \( \alpha = (\alpha_0, \alpha_1^T)^T \), \( \beta = (\beta_0, \beta_1^T, \beta_2^T)^T \), and \( g_\pi(\cdot) \) and \( g_\mu(\cdot) \) are specified link functions. We estimate \( \alpha \) and \( \beta \) using regularized estimators:

\[
\hat{\alpha} = \arg\min_{\alpha} \left\{ -N^{-1} \sum_{i=1}^{N} \ell_\pi(\alpha; X_i, T_i) + p_{\lambda_N}(\alpha_1) \right\} \tag{11}
\]

\[
\hat{\beta} = \arg\min_{\beta} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\mu(\beta; Y_i, X_i, T_i) + p_{\lambda_n}(\beta_{-1}) \right\}, \tag{12}
\]
where \( \ell_\pi(\alpha; X_i, T_i) \) and \( \ell_\mu(\beta; Y_i, X_i, T_i) \) are the log-likelihood contributions for the \( i \)-th observation, and \( p_{\lambda,X}(\cdot) \) and \( p_{\lambda,\mu}(\cdot) \) are penalty functions chosen such that the oracle properties \((\text{Fan and Li} \, 2001)\) hold. Examples of such estimators include the adaptive least absolute shrinkage and selection operator (ALASSO) \((\text{Zou} \, 2006)\) where \( p_{\lambda}(u) = \lambda \sum_{j=1}^p |u_j| / |\tilde{w}_j| \) with initial weights \( \tilde{w}_j \) estimated from ridge regression and tuning parameters are such that \( N\lambda_N \to \infty, \sqrt{N}\lambda_N \to 0, n\lambda_n \to \infty, \) and \( \sqrt{n}\lambda_n \to 0 \). We then calibrate the initial PS estimate \( \pi(x; \hat{\alpha}) \) by the following kernel smoothing estimator:

\[
\hat{\pi}(x; \hat{\alpha}_1, \hat{\beta}_1) = \frac{N^{-1} \sum_{j=1}^N K_{\hat{\lambda}}(\hat{S}_j - \hat{s}) I(T_j = 1)}{N^{-1} \sum_{j=1}^N K_{\hat{\lambda}}(\hat{S}_j - \hat{s})},
\]

where \( \hat{S}_j = (\hat{\alpha}_1, \hat{\beta}_1)^T X_j \) and \( \hat{s} = (\hat{\alpha}_1, \hat{\beta}_1)^T x \) are bivariate scores that represent the covariate in the directions of \( \hat{\alpha}_1 \) and \( \hat{\beta}_1 \), \( K_{\hat{\lambda}}(\cdot) = h^{-2}K(\cdot/h) \), with \( K(\cdot) \) being a bivariate \( q \)-th order kernel with \( q > 2 \) and \( h = O(N^{-\alpha}) \) being a bandwidth for which a suitable choice of \( \alpha \) is discussed below. Finally, we define the proposed SS_{DR} estimator as \( \hat{\mu} = \hat{\mu}_1 - \hat{\mu}_0 \) where:

\[
\hat{\mu}_1 = \left\{ \sum_{i=1}^N \frac{I(T_i = 1)}{\hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}^{-1} \left\{ \sum_{i=1}^N \frac{I(T_i = 1)\hat{Y}_i^\dagger}{\hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}, \tag{13}
\]

and

\[
\hat{\mu}_0 = \left\{ \sum_{i=1}^N \frac{I(T_i = 0)}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}^{-1} \left\{ \sum_{i=1}^N \frac{I(T_i = 0)\hat{Y}_i^\dagger}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}, \tag{14}
\]

with \( \hat{Y}_i^\dagger = g_\xi(\hat{\gamma}^T Z_{\hat{\alpha},i}) \). Here \( \hat{\Delta} \) substitutes the robust imputations based on the PS estimated by the double-index PS \( \hat{Y}^\dagger \) into an IPW estimator weighted also with the double-index PS. Had \( Y \) been fully observed such that \( \hat{Y}_i^\dagger = Y_i \) for \( i = 1, \ldots, N \), an IPW estimator based on the double-index PS is doubly-robust in that it is consistent when a working model for either \( \pi(x) \) in (9) or for \( \mu_k(x) \) in (10) is correctly specified \((\text{Cheng et al.} \, 2017)\). We will show that in the SS setting, using the double-index PS in this robust imputation approach maintains this double-robustness property.

2.5 Asymptotic Robustness and Efficiency Properties of \( \hat{\Delta} \)

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 \( \alpha \in \left( \frac{1-\beta}{2q}, \frac{\beta}{2} \wedge \frac{1}{4} \right) \) and \( n = O(N^{1-\beta}) \),
with \( \beta \in (\frac{1}{q+1}, 1) \), \( \hat{\Delta} \) is doubly-robust so that:

\[
\hat{\Delta} - \Delta = O_p(n^{-1/2})
\] (15)

when either the PS model \( \pi(x; \alpha) \) in (9) or the baseline outcome model \( \mu_k(x; \beta) \) in (10) is correctly specified. To characterize the large sample variability of \( \hat{\Delta} \), we next show it is asymptotically linear and identify its influence function. First define \( \bar{\Delta} = \bar{\mu}_1 - \bar{\mu}_0 \), where:

\[
\bar{\mu}_1 = \mathbb{E} \left\{ \frac{I(T = 1)Y^\dagger}{\pi(X; \alpha_1, \beta_1)} \right\} \quad \text{and} \quad \bar{\mu}_0 = \mathbb{E} \left\{ \frac{I(T = 0)Y^\dagger}{1 - \pi(X; \alpha_1, \beta_1)} \right\},
\]

with \( \pi(x; \alpha_1, \beta_1) = \mathbb{P}(T = 1 \mid \alpha_1^T X = \alpha_1^T \bar{x}, \beta_1^T X = \beta_1^T \bar{x}) \), \( \alpha_1 \) and \( \beta_1 \) as the probability limits of \( \hat{\alpha}_1 \) and \( \hat{\beta}_1 \) regardless of model adequacy, and \( Y^\dagger \) being defined as in (8) except that \( \pi(x) \) is replaced by \( \pi(x; \alpha_1, \beta_1) \). We show in Web Appendix B that, under the same requirements for \( \alpha \) and \( \beta \), the influence function for \( \hat{\Delta} \) is given by \( n^{-1/2}(\hat{\Delta}_k - \Delta_k) = \hat{W}_1 - \hat{W}_0 \), where \( \hat{W}_k = n^{-1/2}(\hat{\mu}_k - \bar{\mu}_k) \) for \( k = 0, 1 \) and:

\[
\hat{W}_k = n^{-1/2} \sum_{i=1}^{n} (v_{\beta_1,k} + u_{pa,\pi,k}) \varphi_{\beta_1,i} + u_{\gamma,k} \varphi_{\gamma,i} + o_p(1),
\] (16)

with \( v_{\beta_1,k} = 0 \) when the PS model \( \pi(x; \alpha) \) is correctly specified and \( u_{pa,\pi,k} = 0 \) when either the PS model \( \pi(x; \alpha) \) or imputation model \( g_\xi(\gamma^T z_\pi) \) without the utility covariate is correctly specified. Here \( \varphi_{\beta_1,i} \) and \( \varphi_{\gamma,i} \) are influence functions for \( \hat{\beta}_1 \) and \( \hat{\gamma} \) such that \( n^{1/2}(\hat{\beta}_1 - \beta_1) = n^{-1/2} \sum_{i=1}^{n} \varphi_{\beta_1,i} + o_p(1) \) and \( n^{1/2}(\hat{\gamma} - \gamma) = n^{-1/2} \sum_{i=1}^{n} \varphi_{\gamma,i} + o_p(1) \). Accordingly, the first term in (16) represents the contribution from estimating \( \beta_1 \) in the baseline outcome model \( \mu_k(x; \beta) \) for the double-index PS appearing in the IPW weight and the utility covariate. The remaining term represents the contribution from estimating \( \gamma \) in the imputation model \( g_\xi(\gamma^T z_\pi) \). The influence function does not include terms associated with the variability in estimating \( \alpha \) 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.

In terms of efficiency, when the PS model \( \pi(x; \alpha) \) is correctly specified, the influence
function in (16) simplifies so that:

\[ n^{1/2}(\hat{\Delta} - \Delta) = n^{-1/2} \sum_{i=1}^{n} (u_{\gamma,1} - u_{\gamma,0})^T \varphi_{\gamma,i} + o_p(1), \]

where:

\[ \varphi_{\gamma,i} = \left[ \mathbb{E}\left\{ Z_{\pi,i}Z_{\pi,i}^Tg_\xi(\gamma^TZ_{\pi,i}) \right\} \right]^{-1} Z_{\pi,i} \{ Y_i - g_\xi(\gamma^TZ_{\pi,i}) \}, \]

for \( g_\xi(u) = \frac{\partial}{\partial u} g_\xi(u) \bigg|_{u=u^*} \). The centering of \( Y_i \) around a model approximation of \( \xi_T(V) \) suggests that \( \hat{\Delta} \) should achieve efficiency gain over complete-case (CC) estimators, which neglect the surrogates \( W \). Let \( \bar{\Delta}^* = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} \) be strictly a functional of the observed data distribution not depending on identification assumptions (2)-(4) as in \( \Delta \). To characterize the efficiency of \( \hat{\Delta} \) in a more full context, we show in Web Appendix C that the semiparametric efficiency bound for \( \bar{\Delta}^* \) under an ideal SS model where the distribution of \((V^T, T)^T\) is known but the conditional distribution of \( Y \) given \((V^T, T)^T\) is unrestricted, with respect to a class of regular parametric submodels subject to mild regularity conditions, is \( \mathbb{E}(\varphi^2_{eff}) \), where:

\[ \varphi_{eff} = U_\pi \{ Y - \xi_T(V) \} \]

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 \((V^T, T)^T\) is unknown. Furthermore, we show in Web Appendix C that \( \hat{\Delta} \) indeed achieves the SS efficiency bound when both the PS and imputation model, \( \pi(x; \alpha) \) and \( g_\xi(\gamma^Tz_\pi) \), are correctly specified so that \( \hat{\Delta} \) is locally semiparametric efficient. In this case, even though the distribution of \((V^T, T)^T\) is actually not known in our setup, the bound under the ideal SS model can still be achieved because \( N \gg n \). The correct specification of \( \mu_k(x; \beta) \) is not required for attaining the efficiency bound, as the bound does not involve \( \mu_k(x) \), but its specification is still important for double-robustness in case \( \pi(x; \alpha) \) is misspecified. The local efficiency of \( \hat{\Delta} \) prompts favorable efficiency compared to CC and other SS estimators that traditionally have sought to be efficient under non-SS nonparametric models where the distribution of \((V^T, T)^T\) is assumed to be unknown. In our SS setting, \( \hat{\Delta} \) gains efficiency over these approaches by
taking advantage of the additional information from the large set of unlabeled data $\mathcal{U}$. Even when the working models are not exactly correct as in practice, we find that $\hat{\Delta}$ still achieves substantial efficiency gains if the models are adequate approximations. In particular, $\mu_k(x; \beta)$ and $g(\gamma^T z_\pi)$ may not be compatible with one another if non-linear models such as logistic regression are used for either working model. Nevertheless, we find in such cases that $\hat{\Delta}$ still attains large efficiency gains over existing estimators when flexible basis functions are used in $g_\xi(\gamma^T z_\pi)$ to more closely approximate $\xi_k(v)$. We offer some more discussion on the efficiency gains of $\hat{\Delta}$ under model misspecification in the Section 5. We next consider inference about $\Delta$ based on $\hat{\Delta}$ through a perturbation resampling procedure.

2.6 Perturbation Resampling

Although the asymptotic variance for $\hat{\Delta}$ is specified through the influence function in (16), a direct estimate is difficult because the influence functions involves complicated functionals of the data distribution. We instead propose a simple perturbation resampling procedure for inference. Let $\mathcal{G} = \{G_i : i = 1, \ldots, N\}$ be non-negative iid random variables with unit mean and variance that are independent of the observed data $\mathcal{D}$. We first obtained perturbed estimators of $\alpha$ and $\beta$:

\[
\hat{\alpha}^* = \arg\min_\alpha \left\{ -n^{-1} \sum_{i=1}^n \ell_\pi(\alpha; X_i, T_i)G_i + p_{\lambda_N}^*(\alpha) \right\}
\]

\[
\hat{\beta}^* = \arg\min_\beta \left\{ -n^{-1} \sum_{i=1}^n \ell_\mu(\beta; Y_i, X_i, T_i)G_i + p_{\lambda_n}^*(\beta_{-1}) \right\},
\]

where $p_{\lambda_N}^*(\cdot)$ and $p_{\lambda_n}^*(\cdot)$ are the corresponding penalties based on the perturbed data if data-adaptive weights are used, such as for adaptive LASSO. This leads to the perturbed double-index PS:

\[
\hat{\pi}_1^*(x; \hat{\alpha}_1^*, \hat{\beta}_1^*) = \frac{\sum_{j=1}^N K_h(\hat{S}_j^* - \hat{s}^*)I(T_j = 1)G_j}{\sum_{j=1}^N K_h(\hat{S}_j^* - \hat{s}^*)G_j},
\]
where \( \hat{S}_j = (\hat{\alpha}_1^*, \hat{\beta}_1^*)^T X_j \) and \( \hat{s}^* = (\hat{\alpha}_1^*, \hat{\beta}_1^*)^T x \) are the perturbed bivariate scores. We then obtain the perturbed estimator for \( \gamma, \hat{\gamma}^* \), as the solution to:

\[
    n^{-1} \sum_{i=1}^{n} Z_{\hat{\pi}^*,i} \left\{ Y_i - g_\xi (\gamma^T Z_{\hat{\pi}^*,i}) \right\} G_i + \lambda_n \gamma_{(-1)} = 0,
\]

where \( \hat{\pi}^* \) specifies that the imputations use utility covariates that plug in \( \hat{\pi}^*(x; \hat{\alpha}_1, \hat{\beta}_1^*) \).

Finally, we calculate the perturbed SS \(_{\text{DR}}\) estimator as \( \hat{\Delta}^* = \hat{\mu}_1^* - \hat{\mu}_0^* \), where:

\[
    \hat{\mu}_1^* = \left\{ \sum_{i=1}^{N} \frac{I(T_i = 1) G_i}{\hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1^*)} \right\}^{-1} \left\{ \sum_{i=1}^{N} \frac{I(T_i = 1) \hat{Y}_i^* G_i}{\hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1^*)} \right\},
\]

and \( \hat{\mu}_0^* = \left\{ \sum_{i=1}^{N} \frac{I(T_i = 0) G_i}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1^*)} \right\}^{-1} \left\{ \sum_{i=0}^{N} \frac{I(T_i = 0) \hat{Y}_i^* G_i}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1^*)} \right\}, \)

with \( \hat{Y}_i^* = g_\xi (\hat{\gamma}^T Z_{\hat{\pi}^*,i}) \). It can be shown based on arguments similar to those in Tian et al. (2007) that the asymptotic distribution of \( n^{1/2}(\hat{\Delta} - \bar{\Delta}) \) coincides with that of \( n^{1/2}(\hat{\Delta}^* - \bar{\Delta}) \mid \mathcal{D} \). In this perturbation scheme, estimation of \( \alpha \), the double-index PS through kernel smoothing, and the final IPW estimators \( \hat{\mu}_1^* \) and \( \hat{\mu}_0^* \) does not technically need to be perturbed as they are estimated based on data from \( \mathcal{U} \), and 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 chose to perturb these steps by default. We approximate the standard error of \( \hat{\Delta} \) based on the empirical standard deviation, or, as a robust alternative, the mean absolute deviation (MAD), of a large number of samples of \( \hat{\Delta}^* \) and construct confidence intervals (CI) based on the empirical percentiles.

3. Simulations

We performed simulations to assess the finite samples bias and relative efficiency (RE) of our proposed estimator (SS \(_{\text{DR}}\)) compared to alternative estimators. In separate simulations we also examined the performance of the perturbation procedure for inference based on SS \(_{\text{DR}}\). For SS \(_{\text{DR}}\), we specified \( h(\cdot) \) in the imputation model in (6) as natural cubic splines with 6 knots.
specified at uniform quantiles. Natural cubic splines were also applied to \( \hat{U}_\pi \) for additional flexibility in the imputation model. Ridge regression with the tuning parameter chosen by cross-validation on the deviance was used for regularization in (7), and adaptive LASSO with initial weights estimated by ridge regression and tuning parameter chosen by minimizing a modified BIC criteria (Minnier et al. 2011) in (11) and (12). A plug-in estimate was used for the bandwidth in the smoothing for the double-index PS (Cheng et al. 2017). Prior to smoothing the components of \( \hat{S} \) 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 \( g_{\xi}(u) = g_\pi(u) = g_\mu(u) = 1/(1 + e^{-u}) \) for the working models in (6) and (10).

For comparison, we considered common CC ATE estimators (Lunceford and Davidian 2004; Kang and Schafer 2007), including the standard IPW estimator (CC_{IPW}), outcome regression estimator (CC_{REG}), and the standard doubly-robust estimator (CC_{DR}). We also considered two estimators that leverage \( \mathcal{Z} \). The first is a naive imputation approach (SS_{Naive}), in which \( Y \) is imputed using a logistic regression of \( Y \) on \( V \) and \( T \), and the imputations are plugged into a standard IPW estimator. The second is adapted from a principled estimator for pretest-posttest randomized studies with \( Y \) missing at random (SS_{PrePost}) (Davidian et al. 2005). We modified the estimator by replacing instances of the randomization probability by PS estimates \( \pi(X_i; \hat{\alpha}) \). SS_{PrePost} is also doubly-robust so that it is consistent for \( \Delta \) when either the PS model \( \pi(x; \alpha) \) or baseline outcome model \( \mu_k(x; \beta) \) is correctly specified, providing another approach to leverage \( \mathcal{Z} \) without requiring correct specification of an imputation model. However, SS_{PrePost} is constructed to achieve the efficiency bound in a model where the distribution of \( (V^T, T)^T \) is unknown and \( Y \) is missing at random with the missingness proportion bounded away from 0. This bound differs from that of the SS model we consider,
and the RE simulations correspondingly show that SS_{dr} is more efficient under a SS setup. For all reference methods, the same logistic regression models with main effects only were used for the underlying requisite PS, baseline outcome, and imputation models.

To mimic the EMR data, we considered the case with $Y$ as binary and $\mathbf{W}$ as count variables. In all scenarios, data were generated according to $\mathbf{X} \sim N\{\mathbf{0}, \sigma_x^2(1 - \rho_x)\mathbf{I} + \sigma_x^2\rho_x\}$, $T | \mathbf{X} \sim \text{Bernoulli}\{\pi(\mathbf{X})\}$, $Y | \mathbf{X}, T \sim \text{Bernoulli}\{\mu_T(\mathbf{X})\}$, and $\mathbf{W} = [\Gamma(1, \mathbf{X}^T, T, Y)^T + \epsilon]$, where $\epsilon \sim N\{0, \sigma_w^2(1 - \rho_w)\mathbf{I} + \sigma_w^2\rho_w\}$ and $\lfloor \cdot \rfloor$ is the floor function. 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 $\Gamma_{5 \times 13} = (0_{5 \times 1}, .11_{5 \times 5}, -.11_{5 \times 5}, .11_{5 \times 1}, (5, 5, 2.5, 0, 0)^T)$. We varied the simulations over different model specifications and sample sizes. The imputation model was misspecified across all settings, whereas three scenarios were considered for the baseline outcome model $\mu_k(\mathbf{x}; \beta)$ and the PS model $\pi(\mathbf{x}; \alpha)$: (1) both models are correctly specified, (2) the PS model is correctly specified but outcome model is misspecified, and (3) the outcome model is correctly specified but the PS model is misspecified. The true $\pi(\mathbf{x})$ and $\mu_k(\mathbf{x})$ for these three settings were specified as:

1. **Both correct:** $\mu_k(\mathbf{x}) = g_\mu(\beta_0 + \beta_1^T \mathbf{x} + \beta_2 k)$, $\pi(\mathbf{x}) = g_\pi(\alpha_0 + \alpha_1^T \mathbf{x})$

2. **Misspecified $\mu$:** $\mu_k(\mathbf{x}) = g_\mu(\beta_0 + \beta_1^T \mathbf{x} + \beta_2 k)$, $\pi(\mathbf{x}) = g_\pi(\alpha_0 + \alpha_1^T \mathbf{x})$

3. **Misspecified $\pi$:** $\mu_k(\mathbf{x}) = g_\mu(\beta_0 + \beta_1^T \mathbf{x} + \beta_2 k)$, $\pi(\mathbf{x}) = g_\pi(\alpha_0 + \alpha_1^T \mathbf{x} + 1)$

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

- $\alpha_0 = -3$, $\alpha_1 = .351_{1 \times 10}$, $\beta_0 = -.65$, $\beta_1 = (11_{1 \times 3} , .51_{1 \times 3} , -1.15 , -11_{1 \times 3})^T$
- $\alpha_{1[1]} = .5(0, .35, 0, .35, 0, .35, 0, .35, 0, .35)^T$, $\alpha_{1[2]} = (.35, 0, .35, 0, .35, 0, .35, 0, .35, 0)^T$
- $\beta_{1[1]} = .5(1, 0, 1, 0, .5, 0, -.5, 0, -1, 0)^T$, $\beta_{1[2]} = (0, .5, 0, .5, 0, .5, 0, .5, 0, .5)^T$.

We considered sample sizes of (A) $n = 100$ with $N = 1,112$, (B) $n = 250$ with $N = 5,000$, and (C) $n = 500$ with $N = 125,000$. The results in each scenario are summarized from 1,000 simulated datasets.
Table 1 presents the bias and root-mean square error (RMSE) across misspecification scenarios. SS_{DR}, SS_{PrePost}, and CC_{DR} exhibits low bias that diminishes to zero as sample size increased under all three scenarios, demonstrating their double-robustness. In contrast, CC_{REG} and CC_{IPW} exhibit substantial bias when working models $\mu_k(x; \beta)$ and $\pi(x; \alpha)$ are misspecified, respectively. SS_{Naive} requires correct specification of both the imputation model $g_{\xi}(\gamma^T z_\pi)$ and PS model $\pi(x; \alpha)$ for consistency but shows negligible bias when both $\mu_k(x; \beta)$ and $\pi(x; \alpha)$ are correctly specified, as the logistic regression imputation model likely provides an adequate approximation under the given data generating process. Still, it incurs substantial bias if either $\pi(x; \alpha)$ or $\mu_k(x; \beta)$ are misspecified, where $g_{\xi}(\gamma^T z_\pi)$ becomes further misspecified in the later case.

[Table 1 about here.]

Figure 1 presents the RE of various estimators relative to CC_{DR} across different scenarios. In the small sample case where $n = 100$ and $N = 1,112$, SS_{DR} is much more efficient than both the CC and other SS estimators, regardless of the specification scenario. It gains over the other SS estimators since its asymptotic variance approximates the SS efficiency bound and through the use of regularization to estimate nuisance parameters. The efficiency gain is most prominent under misspecification of $\mu_k(x; \beta)$, which may be driven by the lack of impact on the influence function by $\mu_k(x; \beta)$ when $\pi(x; \alpha)$ is correctly specified. In the large sample case where $n = 500$ and $N = 125,000$, SS_{DR} is still uniformly most efficient, but the gains are somewhat less pronounced. This is expected at least in part from the reduced role of regularization in large samples. Though SS_{Naive} may appear to be competitive in large samples if $\pi(x; \alpha)$ is correctly specified, it may suffer drastic efficiency loss from misspecification of its imputation model under other data generating processes.

[Figure 1 about here.]

To implement the perturbation procedure, we used the weights $G_i \sim 4 \times Beta(.5, 1.5)$ and
1,000 sets of $G$ for SE and CI estimation. We considered evaluating the perturbations only in the scenario where both $\mu_k(x; \beta)$ and $\pi(x; \alpha)$ were correctly specified models. The results are presented in Table 2. In small samples, SEs estimated by the standard deviation tended to over-estimate due to the presence of outlying perturbed estimates, while SEs estimated by MAD tended to under-estimate. In larger samples, the SE estimation improves. The coverage probabilities are close to nominal levels but slightly under-cover in the sample sizes considered. In other simulations not reported we found that perturbation with weights sampled from a multinomial distribution of size $N$ and $N$ categories with equal probabilities, which effectively implements the bootstrap, to exhibit improved coverage probabilities. However, justifying the bootstrap may be more involved due to the correlated weights.

[Table 2 about here.]

4. EMR Data Application

We applied SS$_{DR}$ and the alternative estimators to compare the rates of treatment response to two biologic agents for treating patients with inflammatory bowel disease (IBD) using the EMRs of two large metropolitan academic medical centers. 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 clinical information needed to ascertain meaningful outcomes and covariates for adjustment. EMRs are thus uniquely positioned to provide evidence on the comparative effectiveness of these two therapies.

The data we considered consisted of $N = 1,243$ total IBD patients, including 200 who initi-
ated 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. non-responder) within one year of treatment initiation. We included 12 baseline covariates to adjust for confounding in \( \mathbf{X} \), including demographics, comorbidities, prior utilization, and inflammation biomarker levels. We also selected 35 post-treatment surrogates for \( \mathbf{W} \), comprising of counts of NLP mentions of clinically relevant terms (e.g. “bleeding”, “fistula”, “tenesmus”) within one year of initiation. The transformation \( u \mapsto \log(1 + u) \) was applied to all count variables in \( \mathbf{V} \) 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 \( \text{SS}_{\text{DR}} \), 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 well-known equivalence between significance tests and confidence sets \( \text{Liu and Singh, 1997, Davison et al, 2003} \). 

As shown in Table 3, the point estimates of most estimators agreed that patients receiving ADA experienced lower rates of treatment response, after adjustment for confounding. \( \text{SS}_{\text{DR}} \) 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, suggesting that patients receiving IFX experience a modest benefit in the rate of response.

5. Discussion

The lack of direct observation on gold-standard outcomes of interest makes it challenging to perform CER using EMR data. Under a SS setting where the true outcome is labeled for only a small subset of patients, we developed an efficient and robust estimator for the ATE
that addresses the missingness in the outcome and confounding bias. The estimator adopts an imputation approach to leverage surrogate data from $\mathcal{Z}$ to improve efficiency. It is not only robust to misspecification of the imputation model but also possesses the traditional doubly-robust property, requiring only correct specification of either a PS or baseline outcome model to be consistent. We showed that it is locally semiparametric efficient under an ideal SS semiparametric model and demonstrated through simulations that it is more efficient than available CC and alternative SS estimators, even under misspecification of working models.

The efficiency gain over CC estimators under a nonparametric model is not obvious if the imputation model is badly misspecified. In a favorable scenario for CC estimators, at distributions of the data for which $\pi(x; \alpha)$ and $\mu_k(x; \beta)$ are correctly specified so that efficient CC estimators under the nonparametric model are available and yet $g_\xi(\gamma^Tz_\pi)$ is misspecified, it can be shown by comparing the asymptotic variances that $\hat{\Delta}$ will still be more efficient if $\mathbb{E}[U_\pi^2\{Y - g_\xi(\gamma^Tz_\pi)\}\{g_\xi(\gamma^Tz_\pi) - \mu_T(X; \beta)\}] = 0$. This condition can be guaranteed, for example, if linear link functions are used for $g_\xi(\cdot)$ and $g_\mu(\cdot)$, and each model includes interactions between $U_\pi^2\hat{\pi}_i$ and its linear predictor. Another potential approach is to include $U_\pi^2\hat{\pi}_i\mu_{T_i}(X_i; \hat{\beta})$ among the covariates of the imputation model and to introduce a more general parameterization of the imputation model given by $g_{\xi, \rho}(\gamma^Tz_\pi)$ such that $\rho$ is a parameter enforcing $\mathbb{E}[U_\pi^2\{Y - g_{\xi, \rho}(\gamma^Tz_\pi)\}g_{\xi, \rho}(\gamma^Tz_\pi)] = 0$ at some $\rho = \rho_0$.

We have assumed that the true outcomes $Y$ are labeled completely at random, which is usually reasonable since researchers can control the labeling. 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, in which case the labeling decision may not have been random. 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, 1995; Robins and Rotnitzky, 1995) to the estimating equation when estimating $\gamma$ in (7). The working
imputation model $g_{x}(\gamma^{T}z_{x})$ in (6) was specified as a main effects model for simplicity, but interactions between $h(V)$ and $T$ could be directly included without difficulty. Interaction effects between $X$ and $T$ can also be accommodated in the baseline outcome model $\mu_{k}(x;\beta)$ in (10) when estimating the double-index PS by estimating the double-index PS separately by treatment groups (Cheng et al., 2017). In the case where $W$ is high dimensional, 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 $n$.

References

An, H. and Little, R. J. (2008). Robust model-based inference for incomplete data via penalized spline propensity prediction. *Communications in Statistics: Simulation and Computation* 37, 1718–1731.

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., et al. (2016). Comparative effectiveness of infliximab and adalimumab in crohn’s disease and ulcerative colitis. *Inflammatory bowel diseases* 22, 880–885.

Chen, S. X., Leung, D. H., and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 803–823.

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.

Chen, Y.-H. (2000). A robust imputation method for surrogate outcome data. *Biometrika* pages 711–716.

Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under
double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62, 449–460.

Cheng, D., Chakrabortty, A., Ananthakrishnan, A. N., and Cai, T. (2017). Estimating average treatment effects with a response-informed calibrated propensity score. *arXiv preprint arXiv:1702.01349*.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical science: a review journal of the Institute of Mathematical Statistics* 20, 261.

Davison, A. C., Hinkley, D. V., and Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science* pages 141–157.

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.

Fiks, A. G., Grundmeier, R. W., Margolis, B., Bell, L. M., Steffes, J., Massey, J., and Wasserman, R. C. (2012). Comparative effectiveness research using the electronic medical record: an emerging area of investigation in pediatric primary care. *The Journal of Pediatrics* 160, 719.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* pages 315–331.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24, 726–748.

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R., Bernstam, E. V., Lehmann, H. P., Hripcsak, G., Hartzog, T. H., Cimino, J. J., et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical care* 51, S30.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of
alternative strategies for estimating a population mean from incomplete data. *Statistical science* pages 523–539.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica* pages 949–968.

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.

Manion, F. J., Harris, M. R., Buyuktur, A. G., Clark, P. M., An, L. C., and Hanauer, D. A. (2012). Leveraging ehr data for outcomes and comparative effectiveness research in oncology. *Current oncology reports* **14**, 494–501.

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.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

Robins, J. (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. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

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
Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Tian, L., Cai, T., Goetghebeur, E., and Wei, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* pages 297–311.

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.
In the following, the supporting lemmas of Web Appendix A identify rates of convergence for frequently encountered quantities and also identify the efficient influence function for $\Delta^*$ under a fully nonparametric model. The results in Web Appendix B show that $\hat{\Delta}$ is consistent and asymptotically linear, deriving its influence function. The results in Web Appendix C establish the semiparametric efficiency bound under the SS model and shows that $\hat{\Delta}$ achieves this bound at particular distributions for the data so that it is locally semiparametric efficient. Throughout these Web Appendices we assume that mild regularity conditions required for the double-index PS in Web Appendix A of Cheng et al. (2017) hold.

The following notations facilitate the subsequent derivations. Let $\pi_k(x) = \mathbb{P}(T = k \mid X = x)$ for $k = 0, 1$. Let $\pi_k(x; \alpha_1, \beta_1) = \mathbb{P}(T = k \mid \alpha_1^T X = \beta_1^T x)$, $\pi_k(x; \alpha) = \pi(x; \alpha)^k \{1 - \pi(x; \alpha)\}^{1-k}$, and $\hat{\pi}_k(x; \alpha_1, \beta_1) = \hat{\pi}(x; \alpha_1, \beta_1)^k \{1 - \hat{\pi}_k(x; \alpha_1, \beta_1)\}^{1-k}$ for given $\alpha_1, \beta_1 \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^{p+1}$ and $k = 0, 1$. Moreover, let $\bar{\vartheta} = (\hat{\alpha}_1^T, \hat{\beta}_1^T)^T$, $\tilde{\vartheta} = (\hat{\alpha}_1^T, \hat{\beta}_1^T)^T$, $\pi_k(x; \tilde{\vartheta}) = \pi_k(x; \alpha_1, \beta_1)$, $\hat{\pi}_k(x; \bar{\vartheta}) = \hat{\pi}_k(x; \alpha_1, \beta_1)$, and $\hat{\pi}_k(x; \bar{\vartheta}) = \hat{\pi}_k(x; \hat{\alpha}_1, \hat{\beta}_1)$. Let the working imputation model be denoted by $\xi_T(V, \gamma, \pi) = g_k\{\gamma^T(1, h(V)^T, T, U_\pi)^T\}$, where $U_\pi = I(T = 1)/\pi(X) - I(T = 0)/\{1 - \pi(X)\}$, given some PS $\pi$. Let $\omega_k,i = I(T_i = k)/\pi_k(X_i)$, $\tilde{\omega}_k,i = I(T_i = k)/\pi_k(X_i, \tilde{\vartheta})$, $\hat{\omega}_k,i = I(T_i = k)/\hat{\pi}_k(X_i, \bar{\vartheta})$, and $\bar{S} = (\alpha_1, \beta_1)^T X$ with $\bar{S}_i = (\hat{\alpha}_1, \hat{\beta}_1)^T X_i$, for $k = 0, 1$ and $i = 1, \ldots, N$.

**Web Appendix A: Supporting Lemmas**

**Lemma 1:** The rates of uniform convergence for kernel estimators we use are as follows:

\[
\sup_x \left\| \hat{\pi}_k(x; \bar{\vartheta}) - \pi_k(x; \bar{\vartheta}) \right\| = O_p(\tilde{a}_N) \quad \text{(A.1)}
\]

\[
\sup_x \left\| \frac{\partial}{\partial \alpha_1^T} \hat{\pi}_k(x; \bar{\vartheta}) - \frac{\partial}{\partial \alpha_1^T} \pi_k(x; \bar{\vartheta}) \right\| = O_p(\tilde{b}_N) \quad \text{(A.2)}
\]

\[
\sup_x \left\| \frac{\partial}{\partial \beta_1^T} \hat{\pi}_k(x; \bar{\vartheta}) - \frac{\partial}{\partial \beta_1^T} \pi_k(x; \bar{\vartheta}) \right\| = O_p(\tilde{b}_N) \quad \text{(A.3)}
\]

\[
\sup_x \left\| \hat{\pi}_k(x; \tilde{\vartheta}) - \pi_k(x; \tilde{\vartheta}) \right\| = O_p(n) \quad \text{(A.4)}
\]
where:
\[ \tilde{a}_N = h^q + \left( \frac{\log N}{Nh^2} \right)^{1/2}, \quad \tilde{b}_N = h^q + \left( \frac{\log N}{Nh^4} \right)^{1/2}, \quad \text{and} \quad a_n = n^{-1/2} + n^{-1/2}\tilde{b}_N + \tilde{a}_N. \]

Proof. The uniform rates for fixed \( \tilde{\alpha}_1 \) and \( \tilde{\beta}_1 \) in the first three equations follow directly from the uniform convergence rates of kernel smoothers and their first derivatives (Hansen 2008). To establish the uniform convergence rate for DiPS, we first note that:
\[
\sup_x \left\| \hat{\pi}_k(x; \hat{\theta}) - \pi_k(x; \hat{\theta}) \right\| \leq \sup_x \left\| \hat{\pi}_k(x; \hat{\theta}) - \hat{\pi}_k(x; \hat{\theta}) \right\| + \sup_x \left\| \hat{\pi}_k(x; \hat{\theta}) - \pi_k(X; \hat{\theta}) \right\|.
\]
The first term on the right-hand side can be written:
\[
\sup_x \left\| \hat{\pi}_k(x; \hat{\theta}) - \hat{\pi}_k(x; \hat{\theta}) \right\| 
\leq \sup_x \left\| \frac{\partial}{\partial \tilde{\alpha}_1} \pi_k(x; \tilde{\alpha}_1, \tilde{\beta}_1)(\tilde{\alpha}_1 - \tilde{\alpha}_1) + \frac{\partial}{\partial \beta_1} \pi_k(x; \tilde{\alpha}_1, \tilde{\beta}_1)(\tilde{\beta}_1 - \beta_1) \right\|
+ \sup_x \left\{ \left( \frac{\partial}{\partial \tilde{\alpha}_1} \hat{\pi}_k(x; \tilde{\alpha}_1, \beta_1) - \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \tilde{\alpha}_1, \beta_1) \right) \right\} (\tilde{\alpha}_1 - \tilde{\alpha}_1)
+ \sup_x \left\{ \left( \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \tilde{\alpha}_1, \beta_1) - \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \tilde{\alpha}_1, \beta_1) \right) \beta_1 - \beta_1 \right\}
+ O_p(\| \tilde{\alpha}_1 - \alpha_1 \|^2 + \| \beta_1 - \beta_1 \|^2 + \| \tilde{\alpha}_1 - \alpha_1 \| \| \beta_1 - \beta_1 \|).
\]
We obtain the desired rate by collecting terms and applying the other rates from above, using that \( \frac{\partial}{\partial \tilde{\alpha}_1} \pi_k(x; \tilde{\alpha}_1, \beta_1) \) and \( \frac{\partial}{\partial \beta_1} \pi_k(x; \tilde{\alpha}_1, \beta_1) \) are continuous in \( x \), and \( x \) lies in a compact covariate space.

**Lemma 2:** Let \( \zeta_i = g(Z_i) \) be some integrable function of \( Z_i = (V_i^T, T_i)^T, \) for \( i = 1, \ldots, N \). Then:
\[
N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \zeta_i = \mathbb{E}(\tilde{\omega}_{k,i} \zeta_i) + O_p(c_n), \quad (A.5)
\]
where \( c_n = \tilde{a}_N + n^{-1/2}N^{-1}h^{-3} \).

Proof. Consider the decomposition:
\[
N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \zeta_i = S_{1,k} + S_{2,k} + S_{3,k},
\]
where:
\[
S_{1,k} = N^{-1} \sum_{i=1}^{N} \bar{\omega}_{i,k} \zeta_{i}, \quad S_{2,k} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{\hat{\pi}_k(X_i; \bar{\vartheta})} - \frac{1}{\pi_k(X_i; \vartheta)} \right\} I(T_i = k) \zeta_i,
\]
and \( S_{3,k} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{\hat{\pi}(X_i; \bar{\vartheta})} - \frac{1}{\pi_k(X_i; \vartheta)} \right\} I(T_i = k) \zeta_i. \)

The second term can be bounded:
\[
|S_{2k}| \leq \sup_x \left\| \hat{\pi}_k(x; \bar{\vartheta}) - \pi_k(x; \vartheta) \right\| N^{-1} \sum_{i=1}^{N} \frac{I(T_i = k) \zeta_i}{\hat{\pi}_k(X_i; \vartheta) \pi_k(X_i; \vartheta)} = O_p(\bar{a}_N).
\]

The third term can be written:
\[
S_{3,k} = N^{-1} \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_1} \frac{I(T_i = k) \zeta_i}{\hat{\pi}_k(X_i; \alpha_1, \beta_1)} (\hat{\alpha}_1 - \alpha_1) + \frac{\partial}{\partial \beta_1} \frac{I(T_i = k) \zeta_i}{\hat{\pi}_k(X_i; \alpha_1, \beta_1)} (\hat{\beta}_1 - \beta_1)
\]
\[
+ O_p \left( \|\hat{\alpha}_1 - \alpha_1\|^2 + \|\hat{\beta}_1 - \beta_1\|^2 + \|\hat{\alpha}_1 - \alpha_1\| \|\hat{\beta}_1 - \beta_1\| \right)
\]
\[
= O_p \left\{ (1 + N^{-1/2} h^{-1} + N^{-1} h^{-3}) n^{-1/2} \right\}
\]
where we use that \( \frac{\partial}{\partial \alpha_1} 1/\hat{\pi}_k(X_i; \alpha_1, \beta_1) \) and \( \frac{\partial}{\partial \alpha_1} 1/\hat{\pi}_k(X_i; \alpha_1, \beta_1) \) are Lipshitz continuous in \( \alpha_1 \) and \( \beta_1 \) for the first equality and the rate deduced from an analogous term in [Cheng et al. (2017)] for the second equality. The desired result follows from collecting the dominant rates.

**Lemma 3:** Let \( Y_i^\dagger = \xi_{T_i}(V_i; \bar{\gamma}, \pi) \) for \( i = 1, \ldots, N \). Then:
\[
\frac{\sqrt{n}}{N} \sum_{i=1}^{N} \bar{\omega}_{k,i}(Y_i^\dagger - \bar{\mu}_k) = O_p(1 + d_n), \quad (A.6)
\]
where \( d_n = \nu_n^{1/2} N^{1/2} h^q + \nu_n^{1/2} N^{-1/2} h^{-2} + N^{-1/2} h^{-1} + N^{-1} h^{-3}. \)

**Proof.** Consider the decomposition:
\[
\frac{\sqrt{n}}{N} \sum_{i=1}^{N} \bar{\omega}_{k,i}(Y_i^\dagger - \bar{\mu}_k) = \hat{W}_{1,1,k} + \hat{W}_{2,1,k} + \hat{W}_{3,1,k},
\]
where:
\[ \tilde{W}_{1,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \tilde{\omega}_{k,i}(Y_i^\dagger - \bar{\mu}_k) \]
\[ \tilde{W}_{2,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left\{ \frac{1}{\tilde{\pi}_k(X_i; \hat{\theta})} - \frac{1}{\pi_k(X_i; \hat{\theta})} \right\} I(T_i = k)(Y_i^\dagger - \bar{\mu}_k) \]
\[ \tilde{W}_{3,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left\{ \frac{1}{\tilde{\pi}_k(X_i; \hat{\theta})} - \frac{1}{\pi_k(X_i; \hat{\theta})} \right\} I(T_i = k)(Y_i^\dagger - \bar{\mu}_k). \]

The first term is a scaled sum of iid centered terms so that:
\[ \tilde{W}_{1,1,k} = \nu_n^{1/2} N^{-1/2} \sum_{i=1}^{N} \tilde{\omega}_{k,i}(Y_i^\dagger - \bar{\mu}_k) = \nu_n^{1/2} O_p(1). \]

The V-statistic arguments similar to Cheng et al. (2017), the second term can be written:
\[ \tilde{W}_{2,1,k} = -\nu_n^{1/2} N^{-1/2} \sum_{i=1}^{N} \mathbb{E}(Y_i^\dagger | \mathbf{S}_i, T_i = k) \left\{ \frac{I(T_i = k)}{\pi_k(X_i; \hat{\theta})} - 1 \right\} + O_p \left\{ \nu_n^{1/2}(h^q + N^{-1/2}h^{-2}) \right\} + O_p(\nu_n^{1/2} N^{1/2}N^{-1/2}h^{-2}) = O_p(\nu_n^{1/2} N^{1/2}h^q + \nu_n^{1/2} N^{-1/2}h^{-2}). \]

The final term can be written:
\[ \tilde{W}_{3,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_1^T} I(T_i = k)(Y_i^\dagger - \bar{\mu}_k) \tilde{\pi}_k(X_i; \alpha_1, \beta_1) + \frac{\partial}{\partial \beta_1^T} I(T_i = k)(Y_i^\dagger - \bar{\mu}_k) (\beta_1 - \bar{\beta}_1) + O_p \left\{ n^{1/2} \left( \| \tilde{\alpha}_1 - \alpha_1 \|^2 + \| \tilde{\beta}_1 - \beta_1 \|^2 + \| \tilde{\alpha}_1 - \alpha_1 \| \| \tilde{\beta}_1 - \beta_1 \| \right) \right\} = O_p(1 + N^{-1/2}h^{-1} + N^{-1}h^{-3}) O_p(1). \]

where we use that \( \frac{\partial}{\partial \alpha_1^T} 1/\tilde{\pi}_k(X_i; \alpha, \beta) \) and \( \frac{\partial}{\partial \alpha_1^T} 1/\tilde{\pi}_k(X_i; \alpha, \beta) \) are Lipshitz continuous in \( \alpha \) and \( \beta \) for the first equality and used the rate deduced from an analogous term from Cheng et al. (2017) for the second equality. The desired result follows from collecting the rates.

**Lemma 4:** Let \( \mathcal{M}_{NP} = \{ f_Y; (y, z) : \text{there exists a } \epsilon_\pi > 0 \text{ such that } \pi_1(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for all } x \text{ where } f_X(x) > 0 \} \) be a nonparametric model for the distribution of \( (Y, Z) \), where \( z = (v^T, t)^T \). Let \( \mathcal{M}_{NP, sub} = \{ f_Y; (y, z; \theta) : \theta \in \Theta \} \) denote a regular parametric submodel of
\( \mathcal{M}_{NP} \), where \( \theta \) is a finite-dimensional parameter and the true density is at \( \theta = \theta^* \). Let \( \mathcal{P}_{NP} \) be the collection of all such regular parametric submodels that satisfy:

1. \( \mathbb{E}_\theta[\mathbb{E}_\theta^*(Y \mid X, T = k)^2] \) is continuous in \( \theta \) at \( \theta = \theta^* \) for \( k = 0, 1 \), where \( \mathbb{E}_\theta(\cdot) \) and \( \mathbb{E}_\theta(\cdot \mid \cdot) \) denote expectation and conditional expectation with respect to \( f(\cdot; \theta) \) and \( f(\cdot \mid \cdot; \theta) \).

2. The score at \( \theta^* \), satisfies
   \[
   S_{Y,W,T,X}(\theta^*) = S_{Y,W,T,X}(\theta^*) + S_{T,X}(\theta^*) + S_X(\theta^*),
   \]
   where \( S_{Y,W,T,X}(\theta^*) \), \( S_{W,T,X}(\theta^*) \), \( S_{T,X}(\theta^*) \) and \( S_X(\theta^*) \) denote the scores in implied parametric submodels for the respective conditional and marginal distributions at \( \theta^* \).

3. \( \frac{\partial}{\partial \theta} \mathbb{E}_\theta^*\{\mathbb{E}_\theta(Y \mid X, T = k)\}_{\theta^*} = \mathbb{E}_\theta^*\{\frac{\partial}{\partial \theta} \mathbb{E}_\theta(Y \mid W, X, T = k)\}_{\theta^*} \mid X, T = k \} \) for \( k = 0, 1 \).

4. \( \mathbb{E}_\theta\{\mathbb{E}_\theta^*(Y \mid W, X, T = k)^2 \mid X, T = k \} \) is continuous in \( \theta \) at \( \theta^* \) for \( k = 0, 1 \).

5. \( \mathbb{E}_\theta(Y^2 \mid W, X, T = k) \) is continuous in \( \theta \) at \( \theta^* \) for \( k = 0, 1 \).

The efficient influence function for \( \tilde{\Delta}^* = \mathbb{E}\{\mathbb{E}(Y \mid X, T = 1) - \mathbb{E}(Y \mid X, T = 0)\} \) in \( \mathcal{M}_{NP} \) with respect to \( \mathcal{P}_{NP} \) is:

\[
\Psi_{\text{eff}} = \mathbb{E}(Y \mid X, T = 1) - \mathbb{E}(Y \mid X, T = 0) + \{\frac{I(T = 1)}{\pi_1(X)} - \frac{I(T = 0)}{\pi_0(X)}\}\{Y - \mathbb{E}(Y \mid X, T)\} - \tilde{\Delta}^*.
\]

The semiparametric efficiency bound for \( \tilde{\Delta}^* \) under \( \mathcal{M}_{NP} \) with respect to \( \mathcal{P}_{NP} \) is \( \mathbb{E}(\Psi_{\text{eff}}^2) \).

**Proof.** The derivation of the semiparametric efficiency bound for \( \tilde{\Delta}^* \) under \( \mathcal{M}_{NP} \) directly follows arguments from the well-known works of Robins et al. (1994) and Hahn (1998). It can be shown that the availability of \( W \) in our framework does not alter the bound as \( \mathcal{M}_{NP} \) is a model for the distribution of data where \( Y \) is fully observed. We omit repeating the arguments here for brevity.
WEB APPENDIX B: CONSISTENCY AND ASYMPTOTIC LINEARITY OF \( \hat{\Delta} \)

THEOREM 1: Under the identification assumptions from (2)-(4) of the main text, given a bandwidth of  \( h = O(N^{-\alpha}) \) for \( \frac{1-\beta}{2q} < \alpha < \min\left(\frac{\beta}{2}, \frac{1}{4}\right) \) and \( n = O(N^{1-\beta}) \) with \( \frac{1}{q+1} < \beta < 1 \), \( \hat{\Delta} - \Delta = O_p(n^{-1/2}) \) when either \( \pi_0(x; \alpha) \) or \( \mu_k(x; \beta) \) is correctly specified.

Proof. We first show that \( \hat{\Delta} - \Delta = O_p(n^{-1/2}) \) where \( \hat{\Delta} = \bar{\mu}_1 - \bar{\mu}_0 \). If this can be shown, the limiting estimate is:

\[
\hat{\Delta} = \mathbb{E} \left\{ \frac{I(T = 1)Y}{\pi_1(X; \vartheta)} - \frac{I(T = 0)Y}{\pi_0(X; \vartheta)} \right\} = \Delta,
\]

where the first equality follows from the argument in the main text and the second equality holds when either \( \pi_1(x; \alpha) \) or \( \mu_k(x; \beta) \) are correctly specified \cite{Cheng2017}. It suffices to show that \( \hat{\mu}_k - \bar{\mu}_k = O_p(n^{-1/2}) \), for \( k = 0, 1 \). First note that the normalizing constant can effectively be ignored. By application of Lemma 2 with \( \zeta_i = 1 \), the normalizing constant is:

\[
N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} = 1 + O_p(c_n). \quad \text{(A.7)}
\]

We can now write the standardized mean for the \( k \)-th group as:

\[
n^{1/2}(\bar{\mu}_k - \bar{\mu}_k) = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(\hat{Y}_i^{\dagger} - \bar{\mu}_k) + \left[ \left\{ N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \right\}^{-1} - 1 \right] \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(\hat{Y}_i^{\dagger} - \bar{\mu}_k)
\]

\[
= \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(\hat{Y}_i^{\dagger} - \bar{\mu}_k) + O_p(c_n), \quad \text{(A.8)}
\]

where the last equality follows provided that the main term is \( O_p(1) \). Denote:

\[
\tilde{W}_k = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(\hat{Y}_i^{\dagger} - \bar{\mu}_k).
\]

This can be decomposed as \( \tilde{W}_k = \tilde{W}_{1,k} + \tilde{W}_{2,k} \), where:

\[
\tilde{W}_{1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(Y_i^{\dagger} - \bar{\mu}_k) \quad \text{and} \quad \tilde{W}_{2,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i}(\hat{Y}_i^{\dagger} - Y_i^{\dagger}). \quad \text{(A.9)}
\]

First, focusing on \( \tilde{W}_{2,k} \), we expand \( \hat{Y}_i^{\dagger} = \xi_T(v_i; \bar{\gamma}, \bar{\pi}) \) around \( Y_i^{\dagger} = \xi_T(v_i; \bar{\gamma}, \bar{\pi}) \):

\[
\tilde{W}_{2,k} = \tilde{W}_{2,k}^\pi + \tilde{W}_{2,k}^\gamma + O_p \left\{ \sup_{x} \left\| \hat{\pi}(x; \hat{\vartheta}) - \pi(x; \vartheta) \right\|^2 \right\} + O_p(\|\bar{\gamma} - \gamma\|^2), \quad \text{(A.10)}
\]
\[
\begin{align*}
\tilde{W}_{2,k}^\pi &= \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \hat{\pi}_k(X_i; \hat{\theta}) - \pi_k(X_i; \hat{\theta}) \right\} \quad \text{(A.11)}
\end{align*}
\]

accounts for estimating the DiPS in the imputation with \( \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) = \hat{g}_\xi (\hat{\gamma}^T Z_{\pi,i}) [\hat{\gamma}_3 \{ - \frac{I(T_i=1)}{\pi_k(X_i, \hat{\theta})} \}] \) and \( \hat{g}_\xi(u) = dg_\xi(u)/du \) and

\[
\tilde{W}_{2,k}^\gamma = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_{T_i}(V_i; \hat{\gamma}, \pi) (\hat{\gamma} - \gamma) + O_p(\sup_x \| \hat{\pi}(x; \hat{\theta}) - \pi(x; \hat{\theta}) \| \| \hat{\gamma} - \gamma \|) \quad \text{(A.12)}
\]

accounts for estimating \( \gamma \) in the imputation. We can further decompose \( \tilde{W}_{2,k}^\pi = \tilde{W}_{2,k}^{\text{pa,} \pi} + \tilde{W}_{2,k}^{\text{np,} \pi} \), where:

\[
\begin{align*}
\tilde{W}_{2,k}^{\text{pa,} \pi} &= \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \hat{\pi}_k(X_i; \hat{\theta}) - \pi_k(X_i; \hat{\theta}) \right\} \quad \text{(A.13)}
\end{align*}
\]

The \( \tilde{W}_{2,k}^{\text{pa,} \pi} \) term accounts for the parametric estimation of \( \alpha_1 \) and \( \beta_1 \) in DiPS and is:

\[
\begin{align*}
\tilde{W}_{2,k}^{\text{pa,} \pi} &= \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \frac{\partial}{\partial \alpha_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \hat{\beta}_1)(\hat{\alpha} - \alpha) \\
&+ \frac{\partial}{\partial \beta_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \hat{\beta}_1)(\beta - \hat{\beta}) + O_p(\| \hat{\alpha} - \alpha \|^2 + \| \hat{\beta} - \beta \|^2 + \| \hat{\alpha} - \alpha \| \| \hat{\beta} - \beta \|) \right\} = N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \frac{\partial}{\partial \alpha_1} \pi_k(X_i; \hat{\alpha}_1, \hat{\beta}_1)n^{1/2}(\hat{\alpha} - \alpha) + \frac{\partial}{\partial \beta_1} \pi_k(X_i; \hat{\alpha}_1, \hat{\beta}_1)n^{1/2}(\beta - \hat{\beta}) \right\} \\
&+ O_p(1 + c_n)O_p(b_N)O_p(n^{1/2}) + O_p(1 + c_n)O_p(b_N)O_p(1) = O_p(1 + c_n) + O_p(b_N), \quad \text{(A.14)}
\end{align*}
\]

where the first equality uses the Liptshitz continuity of \( \hat{K}(\cdot) \), the second equality applies Lemma 1 and Lemma 2 taking \( \zeta_i = \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \), and the last equality applies lemma 2 again taking \( \zeta_i = \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \frac{\partial}{\partial \alpha_1} \pi(X_i; \hat{\theta}) \) as well as \( \zeta_i = \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \frac{\partial}{\partial \beta_1} \pi(X_i; \hat{\theta}) \).

The \( \tilde{W}_{2,k}^{\text{np,} \pi} \) term accounts for the nonparametric smoothing in DiPS and can be bounded:

\[
\begin{align*}
\tilde{W}_{2,k}^{\text{np,} \pi} &\leq n^{1/2} \sup_x \| \hat{\pi}_k(x; \hat{\theta}) - \pi_k(x; \hat{\theta}) \| N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \quad \text{(A.15)}
\end{align*}
\]

\[
\begin{align*}
&= O_p(n^{1/2}a_N)O_p(1 + c_n) = O_p(n^{1/2}a_N). \quad \text{(A.16)}
\end{align*}
\]
Returning to (A.11), the following term accounts for the parametric estimation of $\gamma$:

$$\hat{W}_{2,k}^7 = N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_T(V_i; \hat{\gamma}, \pi)n^{1/2}(\hat{\gamma} - \hat{\gamma}) + O_p(1 + c_n)O_p(a_n)O_p(1)$$

$$= O_p(1 + c_n)O_p(1) + O_p(a_n) = O_p(1 + c_n + a_n), \quad (A.17)$$

where the first equality applies Lemma 2 taking $\zeta_i = \frac{\partial}{\partial \gamma} \xi_T(V_i; \hat{\gamma}, \pi)\frac{\partial}{\partial \pi} \xi_T(V_i; \hat{\gamma}, \pi)$ as well as Lemma 1 and the second equality follows from application of Lemma 2 again taking $\zeta_i = \frac{\partial}{\partial \gamma} \xi_T(V_i; \hat{\gamma}, \pi)$. Finally, collecting all the terms, we find:

$$n^{1/2}(\hat{\mu}_k - \bar{\mu}_k) = \tilde{W}_{1,k} + \tilde{W}_{2,k}^p + \tilde{W}_{2,k}^{mp} + \tilde{W}_{2,k}^\gamma + O_p(a_n^2 + n^{-1}) + O_p(c_n)$$

$$= O_p(1 + d_n) + O_p(1 + c_n + \delta_N) + O_p(n^{1/2}a_N) + O_p(1 + c_n + a_n)$$

$$+ O_p(a_n^2 + n^{-1}) + O_p(c_n)$$

$$= O_p(1), \quad (A.18)$$

where the second to last equality applies Lemma 3 and the last equality follows when $h = O(N^{-\alpha})$ for $\frac{1 - \beta}{2q} < \alpha < \min\left(\frac{\beta}{2}, \frac{1}{4}\right)$ and $n = O(N^{1-\beta})$ for $\frac{1}{q+1} < \beta < 1$. This shows that $\hat{\mu}_k - \bar{\mu}_k = O_p(n^{-1/2})$.

**Theorem 2:** Let $\hat{W}_k = n^{1/2}(\hat{\mu}_k - \bar{\mu}_k)$ for $k = 0, 1$ so that $n^{1/2}(\hat{\Delta} - \Delta) = \hat{W}_1 - \hat{W}_0$. Given a bandwidth of $h = O(N^{-\alpha})$ for $\frac{1 - \beta}{2q} < \alpha < \min\left(\frac{\beta}{2}, \frac{1}{4}\right)$ and $n = O(N^{1-\beta})$ with $\frac{1}{q+1} < \beta < 1$, then $\hat{W}_k$ has the influence function expansion of the form:

$$\hat{W}_k = n^{-1/2} \sum_{i=1}^{n} (v_{\beta_1,k} + u_{pa,\pi,k}^T) \varphi_{\beta_1,i} + u_{\gamma,k}^T \varphi_{\gamma,i} + o_p(1), \quad (A.19)$$

where $v_{\beta_1,k} = 0$ when $\pi_1(x; \alpha)$ is correctly specified and $u_{pa,\pi,k} = 0$ when either $\pi_1(x; \alpha)$ or $\xi_k(v; \gamma, \pi)$ without the utility covariate is correctly specified.

**Proof.** As in (A.8) and (A.9) the standardized mean for the $k$-th group can be written:

$$\hat{W}_k = \tilde{W}_k + O_p(c_n) = \tilde{W}_{1,k} + \tilde{W}_{2,k} + O_p(c_n), \quad (A.20)$$
where the first equality follows provided \( \widetilde{W}_k = O_p(1) \). The first term can be written:

\[
\widetilde{W}_{1,k} = \nu_n^{1/2} N^{-1/2} \sum_{i=1}^{N} \omega_{k,i}(Y_i^\dagger - \bar{\mu}_k)
\]

\[
= \nu_n^{1/2} \widetilde{W}_{1,1,k} + \nu_n^{1/2} \widetilde{W}_{2,1,k} + O_p \left\{ \nu_n^{1/2}(h^q + N^{1/2}h^q + N^{-1/2}h^{-2} + N^{1/2}\bar{a}_N^2) \right\}
\]

\[
+ \left\{ v_{\beta_1,k} + O_p \left( N^{-1/2}h^{-1} + N^{-1}h^{-3} \right) \right\} n^{1/2}(\bar{\beta}_1 - \beta_1) + O_p \left\{ \nu_n^{1/2}(1 + N^{-1/2}h^{-1} + N^{-1}h^{-3}) \right\},
\]

where:

\[
\widetilde{W}_{1,1,k} = N^{-1/2} \sum_{i=1}^{N} \omega_{k,i}(Y_i^\dagger - \bar{\mu}_k)
\]

\[\widetilde{W}_{2,1,k} = -N^{-1/2} \sum_{i=1}^{N} E(Y_i^\dagger | \bar{S}, T = k)(\omega_{k,i} - 1) + O_p(h^q + N^{-1/2}h^{-2})\]

\[v_{\beta_1,k} = E \left\{ \bar{K}_h \left( \frac{\bar{S}_j - \bar{S}_i}{h} \right)^\top (1 - \omega_{k,i}) \frac{I(T_i = k)(Y_i^\dagger - \bar{\mu}_k)}{l_k(X_i; \bar{\theta})}(X_j^\dagger - X_i^\dagger)^\top \right\},\]

with \( l_k(x; \bar{\theta}) = \pi_k(x; \bar{\theta})f(x; \bar{\theta}) \) and \( f(x; \bar{\theta}) \) being the joint density of \( \bar{S} \), and \( x^\dagger = (x, 0) \) for any vector \( x \). Here \( v_{\beta_1,k} = O_p(1) \) in general and is \( 0 \) when \( \pi(X; \alpha) \) is correctly specified \((\text{Cheng et al., 2017})\). As in (A.10) and (A.13) of Theorem 1 the second term from (A.20) can be written:

\[\widetilde{W}_{2,k} = \widetilde{W}_{2,p}^{\rho_a} + \widetilde{W}_{2,p}^{\gamma} + \widetilde{W}_{2,k} + O_p(a_n^2 + n^{-1}).\]

Continuing the expansion of \( \widetilde{W}_{2,k}^{\rho_a} \) from (A.14):

\[
\widetilde{W}_{2,k}^{\rho_a} = N^{-1} \sum_{i=1}^{N} \omega_{k,i} \frac{\partial}{\partial \rho} \xi_{T_i}(V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1^T} \pi_k(X_i; \bar{\theta}) n^{1/2}(\tilde{\beta}_1 - \beta_1) + O_p \left\{ (1 + c_n)(\nu_n^{1/2} + \tilde{b}_N) \right\}
\]

\[
= E \left\{ \omega_{k,i} \frac{\partial}{\partial \rho} \xi_{T_i}(V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1^T} \pi_k(X_i; \bar{\theta}) \right\} n^{1/2}(\tilde{\beta}_1 - \beta_1) + O_p \left\{ c_n + (1 + c_n)(\nu_n^{1/2} + \tilde{b}_N) \right\}
\]

\[= n^{-1/2} \sum_{i=1}^{n} u_{\rho_a,\pi,k}^\top \varphi_{\beta_1,i} + O_p(1) + O_p(c_n + \tilde{b}_N),\]

by repeated application of Lemma \( \[ \) where:

\[u_{\rho_a,\pi,k}^\top = E \left\{ \omega_{k,i} \frac{\partial}{\partial \rho} \xi_{T_i}(V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1^T} \pi_k(X_i; \bar{\theta}) \right\}. \tag{A.21}\]

is a constant that is \( 0^\top \) when either \( \pi_k(x; \alpha) \) or \( \xi_k(v; \gamma, \pi) \) without the utility covariate is correctly specified and \( \varphi_{\beta_1,i} \) is the influence function for \( \tilde{\beta} \). For \( \widetilde{W}_{2,k}^{\gamma} \) from (A.15) we have
that \( \tilde{W}_{2,k}^{np, \pi} = O_p(n^{1/2}a_N) \). For \( \tilde{W}_{2,k}^{\gamma} \), continuing from (A.17):

\[
\tilde{W}_{2,k}^{\gamma} = N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_i(V_i; \gamma, \pi)n^{1/2}(\tilde{\gamma} - \gamma) + O_p(a_n)
\]

\[
= \mathbb{E} \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_i(V_i; \gamma, \pi) \right\} n^{1/2}(\tilde{\gamma} - \gamma) + O_p(c_n)O_p(1) + O_p(a_n)
\]

\[
= n^{-1/2} \sum_{i=1}^{n} u_{\gamma,k}^{T} \phi_{\gamma,i} + o_p(1) + O_p(c_n + a_n),
\]

where:

\[
u = \mathbb{E} \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_i(V_i; \gamma, \pi) \right\}
\]

is some constant and \( \phi_{\gamma,i} \) is the influence function for \( \tilde{\gamma} \).

Collecting the results from above, we find:

\[
\hat{W}_k = n^{-1/2} \sum_{i=1}^{n} (v_{\beta_1,k}^{T} + u_{\beta_{pa,\pi,k}}^{T}) \phi_{\beta_1,i} + u_{\gamma,k}^{T} \phi_{\gamma,i} + O_p \left\{ \nu_n^{1/2}(N^{1/2}h^q + N^{-1/2}h^{-2}) \right\}
\]

\[
+ O_p(N^{-1/2}h^{-1} + N^{-1}h^{-3} + b_N + n^{1/2}a_N + c_n + a_n)
\]

\[
= n^{-1/2} \sum_{i=1}^{n} (v_{\beta_1,k}^{T} + u_{\beta_{pa,\pi,k}}^{T}) \phi_{\beta_1,i} + u_{\gamma,k}^{T} \phi_{\gamma,i} + O_p \left\{ \nu_n^{1/2}N^{1/2}h^q + \frac{(logN)^{1/2}}{N^{1/2}h^2} + \nu_n^{1/2} \frac{(logN)^{1/2}}{h} \right\}.
\]

The error terms are \( o_p(1) \) when \( h = O(N^{-\alpha}) \) for \( \frac{1-\beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{4}) \) and \( n = O(N^{1-\beta}) \) for \( \frac{1}{q+1} < \beta < 1 \).

**Web Appendix C: Semiparametric Efficiency**

**Theorem 3:** Let \( \mathcal{M}_{SS} = \{ f_{Y|Z}(y, z) = f_{Y|Z}(y | z)f_{Z}(z) : f_{Z}(z) \text{ is a known density such that there exists a } \epsilon_\pi > 0 \text{ where } \pi_1^*(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for all } x \text{ with } f_{X}(x) > 0 \} \) be an ideal semiparametric semi-supervised model where the distribution of \( Z = (V^T, T)^T \) is known, with \( Z = (V^T, t)^T \) and \( \pi_1^*(x) \) and \( f_X(x) \) being the implied PS and density of \( X \) under \( f_Z(z) \). Let \( \mathcal{M}_{SS,sub} = \{ f_{Y|Z}(y, z; \theta)f_{Z}(z) : \theta \in \Theta \} \) denote a regular parametric submodel of \( \mathcal{M}_{SS} \), where \( \theta \) is a finite-dimensional parameter, and the true density is at \( \theta = \theta^* \). Let \( \mathcal{P}_{SS} \subseteq \mathcal{P}_{NP} \) be the sub-collection of all such regular parametric submodels among \( \mathcal{P}_{NP} \). The efficient influence
function for $\Delta^\star$ in $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ is:

$$\varphi_{eff} = \left\{ \frac{I(T = 1)}{\pi_1(X)} - \frac{I(T = 0)}{\pi_0(X)} \right\} \{Y - \xi_T(V)\},$$

(A.22)

and the semiparametric efficiency bound for $\Delta^\star$ under $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ is $E(\varphi_{eff}^2)$.

Furthermore, the efficiency bound under $\mathcal{M}_{SS}$ is lower than or equal to the efficiency bound under the fully nonparametric model $\mathcal{M}_{NP}$ where the distribution of $Z$ is unknown. That is,

$$E(\varphi_{eff}^2) \leq E(\Psi_{eff}^2).$$

Proof. Let $L_0^2$ denote the Hilbert space of mean 0 square-integrable functions of $(Y, Z)^T$ at the true distribution, with inner product of $v_1, v_2 \in L_0^2$ defined by $\langle v_1(Y, Z), v_2(Y, Z) \rangle = E\{v_1(Y, Z)v_2(Y, Z)\}$. We first show that the tangent space of $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ at the true distribution is the closure of $\{s(Y, Z) \in L_0^2 : E\{s(Y, Z) \mid Z\} = 0\}$, denoted by $\Lambda_{SS}$. Let $s(Y, Z)$ be any bounded element belonging to $\Lambda_{SS}$. Consider the parametric submodel given by $\mathcal{M}_{SS,lt} = \{f_{Y|Z}(y \mid z; \theta) = f_{Y|Z}(y \mid z; \theta)f^*_z(z) : \theta \in (-\varepsilon, \varepsilon)\}$ for some sufficiently small $\varepsilon > 0$, where:

$$f_{Y|Z}(y \mid z; \theta) = f^*_y(y \mid z)\{1 + \theta s(y, z)\}$$

with $f^*_y(y \mid z)$ being the true density. The true density is thus at $\theta = 0$. It can be shown that $\mathcal{M}_{SS,lt}$ and the implied conditional and marginal submodels have proper densities, are regular, and the respective score can be written as the derivative of the log density with respect to $\theta$. It can also be shown through calculations similar to those in analogous arguments for the derivation of Lemma 4 that $\mathcal{M}_{SS,lt}$ belongs in $\mathcal{P}_{SS} \subseteq \mathcal{P}_{NP}$.

The score for $\mathcal{M}_{SS,lt}$ at $\theta = 0$ is $S_{Y|Z}(\theta^*) = s(Y, Z)$, so any bounded element in $\Lambda_{SS}$ belongs in the tangent space of $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ at the true distribution. Since the bounded elements are dense in $\Lambda_{SS}$ and the tangent space is closed, any element $r(Y, W, T, X) \in \Lambda_{SS}$ also belongs in the tangent space. Any element of the tangent space at the true distribution also belongs in $\Lambda_{SS}$ by the regularity of the parametric submodels and
properties of scores. This verifies that the tangent space of $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ at the true distribution is $\Lambda_{SS}$.

We next show that $\Psi_{eff}$ is one influence function for $\tilde{\Delta}^*$ in $\mathcal{M}_{SS}$ at the true distribution with respect to $\mathcal{P}_{SS}$. Recall from Lemma 4 that $\Psi_{eff}$ is the unique influence function for $\tilde{\Delta}^*$ in $\mathcal{M}_{NP}$ with respect to $\mathcal{P}_{NP}$. This means that under any regular parametric submodel $\mathcal{M}_{NP,sub}$ belonging to $\mathcal{P}_{NP}$, $\Psi_{eff}$ satisfies:

$$\frac{\partial}{\partial \theta} \tilde{\Delta}^*(\theta) \bigg|_{\theta^*} = \mathbb{E}_{\theta^*} \{ \Psi_{eff} \mathbf{S}_{Y,WWTXX}(\theta^*) \}.$$ 

Now since $\mathcal{P}_{SS} \subseteq \mathcal{P}_{NP}$, pathwise differentiability of $\tilde{\Delta}^*(\theta)$ at $\theta = \theta^*$ also holds, in particular, under any regular parametric submodel in $\mathcal{P}_{SS}$, with $\Psi_{eff}$ being one influence function.

Finally, to obtain the efficient influence function for $\tilde{\Delta}^*$ in $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ at the true distribution, we identify the orthogonal projection of $\Psi_{eff}$ onto $\Lambda_{SS}$. It can be verified that this projection is $\Pi(\Psi_{eff} | \Lambda_{SS}) = \Psi_{eff} - \mathbb{E}(\Psi_{eff} | \mathbf{Z})$. The efficient influence function in $\mathcal{M}_{SS}$ is thus:

$$\varphi_{eff} = \Pi(\Psi_{eff} | \Lambda_{SS}) = \Psi_{eff} - \mathbb{E}(\Psi_{eff} | \mathbf{Z})$$

$$= (\mathbb{E}(Y | \mathbf{X}, T = 1) + \left[ \frac{I(T = 1)}{\pi_1(\mathbf{X})} \{ Y - \mathbb{E}(Y | \mathbf{X}, T = 1) \} \right])$$

$$- (\mathbb{E}(Y | \mathbf{X}, T = 0) + \left[ \frac{I(T = 0)}{\pi_0(\mathbf{X})} \{ Y - \mathbb{E}(Y | \mathbf{X}, T = 0) \} \right]) - \tilde{\Delta}^*$$

$$- (\mathbb{E}(Y | \mathbf{X}, T = 1) + \left[ \frac{I(T = 1)}{\pi_1(\mathbf{X})} \{ \mathbb{E}(Y | \mathbf{Z}) - \mathbb{E}(Y | \mathbf{X}, T = 1) \} \right])$$

$$+ (\mathbb{E}(Y | \mathbf{X}, T = 0) + \left[ \frac{I(T = 0)}{\pi_0(\mathbf{X})} \{ \mathbb{E}(Y | \mathbf{Z}) - \mathbb{E}(Y | \mathbf{X}, T = 0) \} \right]) + \tilde{\Delta}^*$$

$$= \left\{ \frac{I(T = 1)}{\pi_1(\mathbf{X})} - \frac{I(T = 0)}{\pi_0(\mathbf{X})} \right\} \{ Y - \mathbb{E}(Y | \mathbf{Z}) \}.$$ 

By the Pythagorean theorem, we can verify:

$$\mathbb{E}(\Psi_{eff}^2) = \| \Psi_{eff} \|_{\mathcal{L}^2}^2 = \| \Pi(\Psi_{eff} | \Lambda_{SS}) \|_{\mathcal{L}^2}^2 + \| \Psi_{eff} - \Pi(\Psi_{eff} | \Lambda_{SS}) \|_{\mathcal{L}^2}^2$$

$$\geq \| \Pi(\Psi_{eff} | \Lambda_{SS}) \|_{\mathcal{L}^2}^2$$

$$= \mathbb{E}(\varphi_{eff}^2).$$
**Corollary 1:** Given a bandwidth of $h = O(N^{-\alpha})$ for $\frac{1-\beta}{2q} < \alpha < \min\left(\frac{\beta}{2}, \frac{1}{4}\right)$ and $n = O(N^{1-\beta})$ for $\frac{1}{q+1} < \beta < 1$, when $\pi_1(x; \alpha)$ and $\xi_k(v; \gamma, \pi)$ are correctly specified, then:

$$n^{1/2}(\tilde{\Delta} - \Delta^*) = n^{-1/2} \sum_{i=1}^{n} U_i \{Y_i - \xi_{T_i}(V_i)\} + o_p(1). \quad (A.23)$$

That is, $\tilde{\Delta}$ achieves the semiparametric efficiency bound in the ideal SS semiparametric model where the distribution of $Z$ is known.

**Proof.** From Theorem 2, given an appropriate bandwidth and order of labels, when $\pi_1(x; \alpha)$ is correctly specified:

$$n^{1/2}(\tilde{\Delta} - \Delta^*) = n^{-1/2} \sum_{i=1}^{n} (u_{\gamma,1}^T - u_{\gamma,0}^T) \varphi_{\gamma, i} + o_p(1),$$

where:

$$u_{\gamma,k} \varphi_{\gamma, i} = \mathbb{E} \left\{ \omega_{k,i} \frac{\partial}{\partial \gamma} \xi_{T_i}(V_i; \gamma, \pi) \bigg| _{\gamma = \bar{\gamma}} \right\} \left\{ \frac{\partial}{\partial \eta} \mathbb{E} Z_{\pi,i} \xi_{T_i}(V_i; \gamma, \pi) \bigg| _{\gamma = \bar{\gamma}} \right\}^{-1} Z_{\pi,i} \{Y_i - \xi_{T_i}(V_i; \gamma, \pi)\}$$

$$= \mathbb{E} \left\{ \omega_{k,i} Z_{\pi,i}^T \hat{g}(\gamma^T Z_{\pi,i}) \right\} \mathbb{E} \left\{ Z_{\pi,i} Z_{\pi,i}^T \hat{g}(\gamma^T Z_{\pi,i}) \right\}^{-1} Z_{\pi,i} \{Y_i - \xi_{T_i}(V_i; \gamma, \pi)\}.$$

The first and second equalities assume the usual regularity conditions to obtain the influence function of an estimator that is the solution of an estimating equation and exchange order of differentiation and integration. The influence function for $\tilde{\Delta}$ can then be written as:

$$(u_{\gamma,1}^T - u_{\gamma,0}^T) \varphi_{\gamma, i} = \mathbb{E} \left\{ U_{\pi,i} Z_{\pi,i}^T \hat{g}(\gamma^T Z_{\pi,i}) \right\} \mathbb{E} \left\{ Z_{\pi,i} Z_{\pi,i}^T \hat{g}(\gamma^T Z_{\pi,i}) \right\}^{-1} Z_{\pi,i} \{Y_i - \xi_{T_i}(V_i; \gamma, \pi)\}.$$

The terms involving $Z_{\pi,i}$ is a population weighted least square projection of $U_i$ onto $Z_i$, weighted by $\hat{g}(\gamma^T Z_{\pi,i})$. But since $Z_{\pi,i}$ includes $U_{\pi,i}$, the influence function simplifies:

$$(u_{\gamma,1}^T - u_{\gamma,0}^T) \varphi_{\gamma, i} = U_{\pi,i} \{Y_i - \xi_{T_i}(V_i; \gamma, \pi)\} = U_{\pi,i} \{Y_i - \xi_{T_i}(V_i)\},$$

where the second equality follows when $\xi_k(v; \gamma, \pi)$ is correctly specified.
Figure 1. RE of estimators, defined as the ratio of mean square errors (MSE) relative to CC_{DR}, by model misspecification scenarios over 1,000 simulated datasets. Higher values of RE denotes greater efficiency (lower MSE) relative to CC_{DR}. Higher values of RE denotes greater efficiency (lower MSE) relative to DR-CC.
| Size   | Estimator | Both Correct | Misspecified $\mu$ | Misspecified $\pi$ |
|--------|-----------|--------------|--------------------|--------------------|
|        |           | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| $n = 100, N = 1112$ | CC\textsubscript{IPW} | 0.008 | 0.235 | 0.012 | 0.296 | 0.014 | 0.126 |
|        | CC\textsubscript{REG} | -0.001 | 0.096 | 0.023 | 0.127 | 0.003 | 0.075 |
|        | CC\textsubscript{DR} | -0.001 | 0.118 | 0.014 | 0.190 | 0.002 | 0.075 |
|        | SS\textsubscript{Naive} | -0.004 | 0.104 | 0.012 | 0.115 | 0.017 | 0.068 |
|        | SS\textsubscript{PrePost} | -0.004 | 0.094 | 0.013 | 0.126 | -0.002 | 0.068 |
|        | SS\textsubscript{DR} | -0.006 | 0.077 | -0.005 | 0.083 | 0.000 | 0.048 |
| $n = 500, N = 12500$ | CC\textsubscript{IPW} | -0.001 | 0.115 | 0.004 | 0.117 | 0.020 | 0.060 |
|        | CC\textsubscript{REG} | 0.000 | 0.041 | 0.022 | 0.057 | -0.002 | 0.029 |
|        | CC\textsubscript{DR} | -0.002 | 0.063 | 0.002 | 0.081 | -0.002 | 0.029 |
|        | SS\textsubscript{Naive} | 0.001 | 0.029 | 0.009 | 0.038 | 0.018 | 0.027 |
|        | SS\textsubscript{PrePost} | 0.001 | 0.033 | 0.000 | 0.048 | -0.002 | 0.021 |
|        | SS\textsubscript{DR} | -0.002 | 0.028 | -0.001 | 0.037 | -0.001 | 0.020 |
Table 2

Performance of perturbation resampling for SS_{DR} in 1,000 simulated datasets when both \( \mu_k(x; \beta) \) and \( \pi(x; \alpha) \) are correctly specified. Emp SE: empirical SE of SS_{DR} over simulated datasets; ASE: average of estimated SE based on the standard deviation of perturbed estimates; ASE_{MAD}: average of SE based on MAD of perturbed estimates; RMSE: root-mean square error; Coverage: coverage of 95% percentile CIs.

| Size          | Bias | Emp SE | ASE  | ASE_{MAD} | RMSE | Coverage |
|---------------|------|--------|------|-----------|------|----------|
| \( n = 100, N = 1112 \) | -0.005 | 0.076  | 0.306 | 0.057     | 0.076 | 0.914    |
| \( n = 250, N = 5000 \)  | -0.003 | 0.043  | 0.039 | 0.036     | 0.043 | 0.921    |
Table 3

Point and SE estimates based on MAD for the ATE of ADA vs. IFX, with respect to one-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. 95% CIs are percentile-based CIs from resampling and p-values are for testing $H_0 : \Delta = 0$ based on inverting percentile CIs.

| Estimator  | Estimate | SE  | 95% CI (Pct)       | p-value |
|------------|----------|-----|---------------------|---------|
| $CC_{Naive}$ | 0.014    | 0.099 | (-0.201, 0.177)    | 0.822   |
| $CC_{IPW}$  | -0.227   | 0.325 | (-0.558, 0.164)    | 0.714   |
| $CC_{REG}$  | -0.067   | 0.123 | (-0.273, 0.162)    | 0.732   |
| $CC_{DR}$   | -0.125   | 0.153 | (-0.416, 0.164)    | 0.592   |
| $SS_{Naive}$| -0.051   | 0.088 | (-0.318, 0.065)    | 0.198   |
| $SS_{PrePost}$ | 0.033 | 0.109 | (-0.265, 0.180)    | 0.778   |
| $SS_{DR}$   | -0.067   | 0.036 | (-0.164, -0.002)   | 0.044   |