Statistica Sinica Preprint No: SS-2023-0006	
Title	Identification and Estimation of Treatment Effects on Long-Term Outcomes in Clinical Trials With External Observational Data
Manuscript ID	SS-2023-0006
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0006
Complete List of Authors	Wenjie Hu, Xiao-Hua Zhou and Peng Wu
Corresponding Authors	Peng Wu
E-mails	pengwu@btbu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Statistica Sinica

Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data

Wenjie Hu[a], Xiao-Hua Zhou[a,b] and Peng Wu[c*]

[a]Peking University, [b]Pazhou Lab, [c]Beijing Technology and Business University

Abstract: In biomedical studies, estimating drug effects on chronic diseases requires a long follow-up period, which is difficult to meet in randomized clinical trials (RCTs). The use of a short-term surrogate to replace the long-term outcome for assessing the drug effect relies on stringent assumptions that empirical studies often fail to satisfy. Motivated by a kidney disease study, we investigate the drug effects on long-term outcomes by combining an RCT without observation of long-term outcomes and an observational study in which the long-term outcome is observed but unmeasured confounding may exist. Under a mean exchangeability assumption weaker than the previous literature, we identify the average treatment effects in the RCT and derive the associated efficient influence function and semiparametric efficiency bound. Furthermore, we propose a locally efficient doubly robust estimator and an inverse probability weighted (IPW) estimator. The former attains the semiparametric efficiency bound if all the working models are

∗correspond to: pengwu@btbu.edu.cn.

Xiao-Hua Zhou is supported by National Key R&D Program of China (No. 2021YFF0901400)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

correctly specified, which may be hard to achieve due to the intertwined working models. While the latter has a simpler form and requires much fewer model specifications. The IPW estimator using estimated propensity scores is more efficient than that using true propensity scores and achieves the semiparametric efficient bound in the case of discrete covariates and surrogates with finite support. Both estimators are shown to be consistent and asymptotically normally distributed. Extensive simulations are conducted to evaluate the finite-sample performance of the proposed estimators. We apply the proposed methods to estimate the efficacy of oral hydroxychloroquine on renal failure in a real-world data analysis.

Key words and phrases: Data fusion, Long-term treatment effects, Semiparametric efficiency, Surrogate.

1. Introduction

In biomedical research, randomized clinical trials (RCTs) are the gold standard for drug or therapy evaluation (Cartwright, 2010). However, the high cost of labour and material resources restricts the sample size and the duration of RCTs. Especially for chronic diseases, the important long-term outcomes are difficult to observe during the period of RCTs. As a motivating example, we consider a clinical study on immunoglobulin A nephropathy (IgAN), which is the most prevalent form of primary glomerular disease worldwide (D'amico, 1987). A double-blind randomized clinical trial is con-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

ducted at Peking University First Hospital for six months, to compare the efficacy of additional use of oral hydroxychloroquine (HCQ) with only optimized renin-angiotensin-aldosterone system (RAAS) inhibition, which is a standard therapy for IgAN disease. The outcome of interest is whether the patient will develop renal failure over a period of time. But IgAN is a chronic disease and the long-term outcome is not observed within a six-month experiment period, instead, the researcher collected the percentage change in proteinuria as a surrogate (Liu et al., 2019).

This kind of need to evaluate the long-term effect in clinical trials is pervasive in medical and social science applications and requires new methodologies. If only RCT data is available, to evaluate the effect of a treatment on the long-term outcome, the researchers often choose a short-term surrogate that is strongly predictive of the outcome and can be observed during the RCTs and then report analysis results for the surrogate (Liu et al., 2019). The criteria for choosing short-term surrogates have been studied over the years (Prentice, 1989; Frangakis and Rubin, 2002; Lauritzen et al., 2004; Chen et al., 2007; Ju and Geng, 2010). However, the aim of using a surrogate to replace the outcome of interest is too ambitious (Kallus and Mao, 2020). For example, Chen et al. (2007) raised the surrogate paradox, a phenomenon that treatment has a positive effect on a surrogate that

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

has a positive effect on the outcome of interest, but the treatment has a negative effect on the outcome of interest. Stringent unverifiable assumptions are made to avoid the surrogate paradox (Chen et al., 2007). Thus it is important to propose more flexible methods that rely on less stringent assumptions to estimate the treatment effects on the long-term outcomes. Besides RCT data, hospitals usually have a large amount of observational data containing long-term outcomes. Nevertheless, the existence of unmeasured confounders is unavoidable in an observational study (Kallus and Zhou, 2018; Ding et al., 2022), which will impede valid inference about the target quantity such as the efficacy of a newly developed therapy. This article aims to identify the drug effect on long-term outcomes in RCT by combining an RCT dataset and an observational dataset.

In this paper, we mainly make the following three contributions. First, under a mean exchangeability assumption, we elaborate on the identifiability of the treatment effect on long-term outcomes in RCT by combining an RCT without observation of long-term outcome and an observational study in which the long-term outcome is observed but unmeasured confounding may exist. We show that the identifiability assumptions adopted in this article are weaker than those of existing methods. Second, we derive the efficient influence function and the semiparametric efficiency bound for the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

target parameter. Third, we propose a locally efficient doubly robust (DR) estimator and an inverse probability weighted (IPW) estimator, and show their large sample properties. The proposed DR estimator is locally efficient in the sense that it attains the semiparametric efficiency bound if all the working models are correctly specified. However, the proposed DR estimator relies on the estimations of multiple complex nuisance parameters contained in the efficient influence function, its efficiency may degrade significantly if some of the nuisance parametric models are misspecified. This is not a problem unique to our method, existing approaches have similar problems (see Athey et al., 2019, 2020; Kallus and Mao, 2020; Chen and Ritzwoller, 2023). To ease this problem, we further propose a simpler IPW estimator that relies on much fewer nuisance parameters and shows that using an estimated propensity score will lead to better performance even if we know the true propensity score, which is common for RCTs (Robins et al., 1994; Hirano et al., 2003). In addition, we show that the IPW estimator with estimated propensity scores achieves the semiparametric efficient bound in the case of discrete covariates and surrogates with finite support. Both the proposed doubly robust and IPW estimators are shown to be consistent and asymptotically normally distributed, and the associated variance can be estimated. Extensive simulations are conducted to evaluate

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

the performance of the proposed estimators.

The idea of leveraging external data to help identify and improve efficiency has gained much attention in the field of causal inference (Bareinboim and Pearl, 2016; Hünermund and Bareinboim, 2024; Kallus et al., 2018; Yang, Zeng, and Wang, Yang et al.; Yang et al., 2022; Yang and Ding, 2020; Li et al., 2021; Wu et al., 2022, 2023). Our method is closely related to the recently proposed methods of studying the long-term treatment effect. Athey et al. (2019) considered identifying the long-term causal effect in RCT data in a setting where the long-term outcome is not observed in RCT data and the treatment variable is missing in observational data. They briefly discussed estimation methods for the average treatment effect in RCT data, without showing the large sample properties of the proposed estimator. Different from their setting, we assume that treatment variables are observed in the observational study and allow treatment to have a direct effect on outcomes. We identify the same parameter under weaker assumptions than those in Athey et al. (2019), derives the semiparametric efficiency bound, proposes two new estimators, and shows their asymptotic properties. In addition, Kallus and Mao (2020) considered the efficiency gain of estimating the causal effect on a long-term outcome by using an observed short-term surrogate when the long-term outcome is missing at

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

random. However, the missing at random assumption is less plausible in the combined RCT and observational data.

Several works have considered a causal parameter similar to ours, which is the long-term causal effects in observational data (Athey et al., 2020; Ghassami et al., 2022; Imbens et al., 2022; Chen and Ritzwoller, 2023), arguing that this quantity may have better generalizability. But in practice, there are occasions when RCT is a representative sample of the target population, for example, Li et al. (2021); Athey et al. (2020) consider the average treatment effect in the RCT data for a new drug or a policy. Besides, real-world RCTs such as "pragmatic randomized clinical trial" can contain samples reflective of real-world population Gamerman et al. (2019) and become more popular in recent years. Furthermore, in empirical medical data analysis, the analysis results based on RCT data are more credible and are more easily accepted by regulators such as FDA. Therefore, we choose the long-term causal effects in the RCT as the target parameter.

The rest of the article is organized as follows. In Section 2, we describe the setting of the problem interested and give the identifiability assumptions, and compare them with the existing approaches. Section 3 shows the semiparametric efficiency bound for the target parameter. Section 4 proposes two new estimators and presents their large sample properties. In

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Section 5, extensive simulations are performed to evaluate the finite sample behaviors of the proposed methods. Section 6 illustrates our approaches with an empirical example. A brief discussion is concluded in Section 7.

2. Causal Parameter and Identifiability

2.1 Study design and causal parameter

When combining datasets from different sources, sampling mechanisms of the multiple datasets are crucial for statistical inference. There are mainly two ways to view the study design of the RCT data and observational data: nested design and non-nested design (Colnet et al., 2024; Dahabreh et al., 2021). In this paper, we adopt the non-nested design where the sampling mechanisms of the RCT are independent of the observational data. Suppose that there exists an underlying population for the patients, and two subpopulations with different distributions. The RCT data and observational data are simple random samples from two corresponding subpopulations, where the sampling probabilities for the two subpopulations are unknown. With the observed data, the distributions of the two subpopulations can be identified, while the underlying population is not because of the unknown sampling probabilities. And the overall population for the observed data consists of samples from two subpopulations. A more detailed discussion of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

the related study design can be found in Li et al. (2021).

Now we introduce the observed data structure in our problem. Let T denote the indicator for binary treatment, with $T = 1$ or 0 the treated or control group, X denotes the observed pre-treatment covariates, Y denotes the long-term outcome of interest, and S denote the short-term surrogates (e.g., intermediate outcomes) that are highly informative about the outcome Y and measured after the treatment T. Under the potential outcome framework (Rubin, 1974; Neyman, 1990), let $\{S(1), Y(1)\}$ and $\{S(0), Y(0)\}$ be the potential outcomes with and without treatment respectively. The observed surrogate S and outcome Y are the potential outcomes corresponding to the treatment received by the consistency assumption, i.e. $S = S(T)$ and $Y = Y(T)$. Suppose that we have available two data sources: an RCT dataset $\{(T_i, X_i, S_i) : i = 1, ..., n_1\}$ consists of independent and identically distributed (i.i.d.) sample of n_1 observations, and an i.i.d. observational dataset $\{(T_i, X_i, S_i, Y_i) : i = n_1 + 1, ..., n_1 + n_0\}$ contains n_0 observations. Therefore, the observed data has sample size $n = n_0 + n_1$. Denote $G_i \in \{0, 1\}$ as the indicator of the data sources, where $G_i = 1$ represents that unit i belongs to RCT data and $G_i = 0$ represents that unit i belongs to observational data. The limit of n_1/n as $n \to \infty$ tends to a positive constant $q = \text{pr}(G = 1)$, which represents the proportion of the RCT data

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

in the observed data population. The parameter of interest is the average treatment effect in the RCT defined by $\tau = E\{Y(1) - Y(0)|G = 1\}$.

2.2 Assumptions and identifiability

For identification, Assumptions 1 and 2 are imposed throughout.

Assumption 1 (Internal validity of RCT data) For $t = 0$ or 1,

$$T \perp\!\!\!\perp (Y(t), S(t)) \mid X, G = 1.$$

Assumption 2 (Strict overlap) There exists a constant $0 < \varepsilon < 1/2$, such that

(i) $\varepsilon \leq e(X) := \text{pr}(T = 1 \mid X, G = 1) \leq 1 - \varepsilon$,

(ii) $\varepsilon \leq \text{pr}(T = 1 \mid X, S, G = 0) \leq 1 - \varepsilon$,

(iii) $\varepsilon \leq \text{pr}(G = 0 \mid X = x, S = s)$ for all (x, s) satisfying $\text{pr}(X = x, S = s \mid G = 1) > 0$.

Assumption 1 guarantees that the treatment assignment in RCT is unconfounded and is satisfied in most cases with a carefully designed experiment. Kallus and Mao (2020) uses the assumption $T \perp\!\!\!\perp (Y(t), S(t)) \mid X$, where unconfoundedness holds in the combined data, rather than in RCT data. However, this assumption is less plausible, as the existence of unmea-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

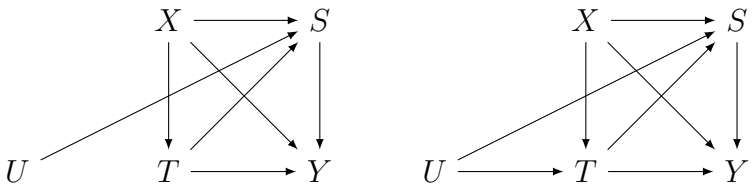
sured confounders is an unavoidable problem for observational data (Kallus et al., 2018). Assumptions 2(i)-(ii) are common in causal inference literature (Rosenbaum and Rubin, 1983; Tsiatis, 2006; Imbens and Rubin, 2015; Hernán and Robins, 2020), which ensure that each unit has the chance to be assigned to each treatment option. Assumption 2(iii) means that each unit in RCT has a positive probability of belonging to the observational data group. This implicitly restricts the support of covariates and surrogates in RCT data should be included in those of observational data, which is necessary to leverage observational data to help identify τ. Besides, Assumption 2(iii) is reasonable in empirical studies because the inclusion rule exerted in RCT will prevent part of the patients from entering the experiment, leading to a smaller support set of X and S. The causal estimand τ is not identified under Assumptions 1 and 2, we further invoke the following mean exchangeability assumption.

Assumption 3 (Mean exchangeability) For $t = 0$ or 1,

$$E(Y(t) \mid X, S(t), T = t, G = 1) = E(Y(t) \mid X, S(t), T = t, G = 0).$$

By consistency, Assumption 3 can be written as $E(Y|X, S, T, G = 1) = E(Y|X, S, T, G = 0)$, an equation only consists of observed data, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006



(a) Observational data (b) Observational data

Figure 1: Typical causal graph when Assumption 3 holds, where U denotes the unmeasured confounder.

the key assumption that enables us to transfer the conditional mean of Y in observational data to RCT data. The mean exchangeability is a weaker version of $G \perp\!\!\!\perp Y(t) \mid X, S(t), T = t$ brought up by Kallus and Mao (2020) and similar assumptions are invoked across various data fusion literature (Li et al., 2021; Miao et al., 2022). Importantly, Assumption 3 allows for the existence of unmeasured confounders between the treatment T and the surrogates S, as the unmeasured confounders have no direct effect on Y, leading to the same conditional expectation of Y between the two datasets. Figures 1(a)-(b) give typical causal graphs when Assumption 3 holds. The following Proposition 1 shows the identifiability of τ.

Proposition 1. *Under Assumptions 1, 2 and 3, τ is identified.*

2.3 Comparison with the identifiability of existing methods

The previous literature adopts stronger identifiability assumptions than Assumptions 1-3. Concretely, Athey et al. (2020) and Chen and Ritzwoller

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

(2023) adopt the following Assumptions 4 and 5 to substitute Assumption 3.

Assumption 4 (Conditional external validity)

$$G_i \perp\!\!\!\perp (Y_i(0), Y_i(1), S_i(0), S_i(1)) \mid X_i$$

Assumption 5 (Latent unconfoundedness) For $t = 0$ or 1,

$$T_i \perp\!\!\!\perp Y_i(t) \mid X_i, S_i(t), G_i = 0$$

Under Assumptions 1, 2, 4, and 5, Athey et al. (2020) and Chen and Ritzwoller (2023) obtain the identifiability of the average treatment effect in observational data, i.e., $E[Y(1) - Y(0)|G = 0]$, and the authors assert that these assumptions are also applicable to identify τ. However, Assumptions 4–5 may be too strong for empirical applications when the focus is the average treatment effect in RCT data. Assumption 4 states that conditioning on X, the distributions of potential outcomes are the same between RCT data and observational data, which implicitly assumes that the distributions of the unmeasured confounders affecting T and S conditional on observed covariates are the same between RCT data and observational data. Compared to Assumptions 4–5, Assumption 3 imposes weaker constraints

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

on the data-generating distribution. In fact, we can show that Assumptions 4–5 are sufficient conditions for Assumption 3 under Assumption 1.

Proposition 2. *Under Assumption 1, Assumption 3 is implied by Assumptions 4 and 5.*

Below we provide an example that satisfies Assumption 3 but violates Assumptions 4-5.

Example 1. Consider the following structural equation models. For RCT data:

$$\text{pr}(T=1)=1/2, S=\alpha_1 X+\alpha_2 U+\tau_S T+\varepsilon_S$$

$$Y=\beta_1 X+\beta_2 S+\tau T+\varepsilon_Y$$

For observational data:

$$\text{pr}(T=1\mid X,U)=\{1+\exp(-\gamma_1 X-\gamma_2 U)\}^{-1}, S=\alpha_1 X+\alpha_2 U+\tau_S T+\varepsilon_S$$

$$Y=\beta_1 X+\beta_2 S+\tau T+\varepsilon_Y$$

where U is an unmeasured variable in both RCT data and observational data, ε_S and ε_Y be the error terms independent of all other variables. If the distribution of $U|G=1$ and $U|G=0$ are different, one can verify that the distribution of $S(t), Y(t), t=0,1$ are different in RCT data and observational data, thus Assumption 4 is violated.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

3. Semiparametric Efficiency Bound

Under the nonparametric model restricted by Assumptions 1–3, we calculate the semiparametric efficiency bound for τ. The following intermediate quantities will appear in the efficient influence function: the selection propensity score $g_t(s,x) = \text{pr}(G=1|S=s,X=x,T=t)$, which quantifies the probability of selection into RCT group for a given surrogate, baseline covariates and treatment; the treatment propensity score for RCT $e(x) = \text{pr}(T=1|X=x,G=1)$; the regression functions $\mu_t(s,x) = E(Y(t)|S(t)=s,X=x,G=1) = E(Y|S=s,X=x,T=t,G=1)$ and $\mu_t(x) = E(Y(t)|X=x,G=1) = E(Y|X=x,T=t,G=1) = E\{\mu_t(S,X)|X=x,T=t,G=1\}$ for $t=0,1$. With these nuisance parameters, Theorem 1 presents the efficient influence function for τ.

Theorem 1 (efficiency bound)**.** *Under Assumptions 1–3, the efficient influence function for τ is given as*

$$\phi = \frac{G}{q}\left\{\frac{T(\mu_1(S,X)-\mu_1(X))}{e(X)} - \frac{(1-T)(\mu_0(S,X)-\mu_0(X))}{1-e(X)} + (\mu_1(X)-\mu_0(X)) - \tau\right\}$$
$$+ \frac{1-G}{q}\left\{\frac{g_1(S,X)T\{Y-\mu_1(S,X)\}}{e(X)\{1-g_1(S,X)\}} - \frac{g_0(S,X)(1-T)\{Y-\mu_0(S,X)\}}{\{1-e(X)\}\{1-g_0(S,X)\}}\right\},$$

where $q = p(G=1)$. The semiparametric efficiency bound is $E(\phi^2)$. In

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

addition,

(i) the efficiency bound remains the same no matter whether the propensity score $e(X)$ is known or not.

(ii) ϕ is the unique influence function in the nonparametric model class that is only restricted by Assumptions 1–3.

Theorem 1 shows that for any regular and asymptotic linear estimator, its asymptotic variance is no smaller than the efficiency bound $E(\phi^2)$. Chen and Ritzwoller (2023) obtains the efficient influence function for average treatment effect for the long-term outcome in observational data under Assumptions 1, 2, 4, and 5. Here our focus is the average treatment effect in RCT data, and the efficient influence function is derived under weaker assumptions. Theorem 1(i) shows that the propensity score is ancillary for the estimation of τ, that is, the knowledge of $e(x)$ does not decrease the efficiency bound of τ. Similar conclusions that knowing some nuisance parameters will not change the efficiency bound of the target parameter are made in Hahn (1998) and Chen and Ritzwoller (2023). The uniqueness of the influence function in Theorem 1(ii) means that any regular and asymptotic linear estimators for τ in the nonparametric model have the same influence function and thus the same asymptotic distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

4. Estimation

4.1 Efficient doubly robust estimator

Theorem 1 motivates an estimator that can achieve the semiparametric efficiency bound. Concretely, generalized linear models are specified for the nuisance parameters in ϕ, including $\mu_t(s, x; \alpha_t)$ and $\mu_t(x; \beta_t)$ for $t = 0$ or 1, $e(x; \gamma)$, and $g_t(s, x; \eta_t)$. Let $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{\gamma}$, and $\hat{\eta}_t$ denote the maximum likelihood estimators of α_t, β_t, γ, η_t, respectively.

The estimation of $e(x; \gamma)$ is trivial. However, particular care is needed when estimating $g_t(s, x; \eta_t)$, $\mu_t(s, x; \alpha_t)$ and $\mu_t(x; \beta_t)$. First, $g_t(s, x; \eta_t)$ should be estimated based on the combined samples of both RCT and observational data; Second, we cannot estimate $\mu_t(s, x; \alpha_t)$ directly due to the missingness of Y in RCT data. Owing to Assumption 3, $\mu_t(s, x; \hat{\alpha}_t)$ can be obtained by regressing Y on (X, S) in observational data with $T = t$, then we calculate their predicted values in RCT data ; Finally, $\mu_t(x; \hat{\beta}_t)$ can be derived by conducting a linear regression of $\mu_t(s, x; \hat{\alpha}_t)$ on X in the RCT sample. It should be noted that we can't estimate $\mu_t(x; \beta_t)$ by directly regressing Y on X with observational data, since $E[Y|X, T = t, G = 1]$ may not equal to $E[Y|X, T = t, G = 0]$ under Assumptions 1–3. With these fitted nuisance

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

parameters, the efficient doubly robust estimator is given as

$$\hat{\tau}_{dr} = \hat{E}\Bigg[\frac{G}{\hat{q}}\Bigg\{\frac{T(\mu_1(S,X;\hat{\alpha}_1)-\mu_1(X;\hat{\beta}_1))}{e(X;\hat{\gamma})} - \frac{(1-T)(\mu_0(S,X;\hat{\alpha}_0)-\mu_0(X;\hat{\beta}_0))}{1-e(X;\hat{\gamma})} + \mu_1(X;\hat{\beta}_1) - \mu_0(X;\hat{\beta}_0)\Bigg\} + \frac{1-G}{\hat{q}}\Bigg\{\frac{g_1(S,X;\hat{\eta}_1)T\{Y-\mu_1(S,X)\}}{e(X;\hat{\gamma})\{1-g_1(S,X;\hat{\eta}_1)\}} - \frac{g_0(S,X;\hat{\eta}_0)(1-T)\{Y-\mu_0(S,X)\}}{\{1-e(X;\hat{\gamma})\}\{1-g_0(S,X;\hat{\eta}_0)\}}\Bigg\}\Bigg],$$

where $\hat{E}(\cdot)$ denotes the sample average of all data throughout, $\hat{q} = n_1/(n_1+n_0)$. The large sample properties of $\hat{\tau}_{dr}$ are presented in the following Theorem 2.

Theorem 2. *Under Assumptions 1–3 and regularity conditions described in theorems 2.6 and 3.4 of Newey and McFadden (1994), the estimator $\hat{\tau}_{dr}$ is consistent and asymptotically normal if either*

(i) the outcome model $\mu_t(S,X;\alpha_t)$ and $\mu_t(X;\beta_t)$ for $t=0,1$ are correctly specified, or

(ii) the outcome model $\mu_t(S,X;\alpha_t)$ and the propensity score model $e(X;\gamma)$ are correctly specified.

In addition, $\hat{\tau}_{dr}$ is locally efficient, i.e., it attains the semiparametric efficiency bound $E(\phi^2)$ when all the working models are correctly specified.

Theorem 2 indicates that the consistency of $\hat{\tau}_{dr}$ relies on the correct specifications of $\mu_t(S,X)$ for $t=0,1$, which may not be guaranteed in real data analysis and increases the risk of obtaining biased conclusions. How-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

ever, the consistency of $\hat{\tau}_{dr}$ does not rely on the correct specification of the selection propensity score $g_t(s, x)$, although the asymptotic variance does. Besides, the doubly robust estimator involves many nuisance parameters and some of them are intertwined. For example, the definitions of $\mu_t(S, X)$ and $\mu_t(X)$ imply that $\mu_t(X) = E[\mu_t(S, X)|X, G = 1]$. When a logistic model is specified for $\mu_t(S, X)$, a logistic model for $\mu_t(X)$ can hardly be correctly specified. We found that this is not a problem unique to our method, existing approaches have similar problems (see Athey et al., 2019, 2020; Kallus and Mao, 2020; Chen and Ritzwoller, 2023).

When all the parametric models for the nuisance parameters are correctly specified, the asymptotic variance of $\hat{\tau}_{dr}$ can be naturally estimated by $\hat{E}(\hat{\phi}^2)$, where $\hat{\phi}$ is the plug-in estimator of ϕ. Besides, we can use the bootstrap method to get the asymptotic variance estimation if we cannot ensure the correctness of all model specifications. Concretely, for each bootstrap, we randomly sample n_1 and n_0 samples from RCT and observational data with replacement, respectively. Repeat B times to get B point estimates. Then the sample variance of the B point estimates is the estimate of the asymptotic variance of $\hat{\tau}_{dr}$. In Section 5, our simulation compares these two methods of calculating the asymptotic variance.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

4.2 Inverse probability weighted estimator

The doubly robust estimator has some worrying features. As discussed in Section 4.1, its efficiency relies on the correctness of multiple cumbersome model specifications for the nuisance parameters. When some models are misspecified, the efficiency may degrade and the estimator may have a bias. As a complement, the inverse probability weighted (IPW) estimator that can consistently estimate τ by merely imposing a model specification for $h(X, S, T) = E[Y|X, S, T, G = 1]$, which is given by

$$\hat{\tau}_{ipw} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \frac{T_i \cdot h(X_i, S_i, T_i; \hat{\kappa})}{e(X_i)} - \frac{(1 - T_i) \cdot h(X_i, S_i, T_i; \hat{\kappa})}{1 - e(X_i)} \right\}, \tag{4.1}$$

where $h(X, S, T; \kappa)$ is assumed to be a generalized linear model and $\hat{\kappa}$ is the maximum likelihood estimator of κ based on observational data. Clearly, the IPW estimator has a much simpler form than the doubly robust estimator and thus is more tractable.

Generally, there are two obstacles to applying the IPW estimator in statistical analysis: imprecision and instability when some propensity score values are close to 0 or 1 (Tan, 2007, 2010; Molenberghs et al., 2015; Wu et al., 2021, 2024). Since the propensity score in RCT data is usually known and bounded away from 0 or 1, the problem of instability does not exist in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

our setting. To improve the efficiency of the IPW estimator, we propose using the estimated propensity score, instead of the true propensity score, to construct the IPW estimator. Specifically, we use logistic regression to estimate it, i.e., assuming $e(X_i) = e(X_i;\gamma) = \exp(X_i^T\gamma)/\{1+\exp(X_i^T\gamma)\}$. Let $\hat{\gamma}$ be the maximum likelihood estimator of γ, and define

$$\tilde{\tau}_{ipw} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\frac{T_i\cdot h(X_i,S_i,T_i;\hat{\kappa})}{e(X_i;\hat{\gamma})} - \frac{(1-T_i)\cdot h(X_i,S_i,T_i;\hat{\kappa})}{1-e(X_i;\hat{\gamma})}\right\}. \tag{4.2}$$

Next, we establish the asymptotic properties of $\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$. For ease of exposition hereafter, we let $e_i = e(X_i)$, $\hat{e}_i = e(X_i;\hat{\gamma})$, $\tilde{X}_i = (X_i^T, S_i^T, T_i)^T$, $h_i = h(X_i,S_i,T_i;\kappa)$, $\hat{h}_i = h(X_i,S_i,T_i;\hat{\kappa})$, and $h_i'(\kappa) = \partial h(X_i,S_i,T_i;\kappa)/\partial\kappa$. Denote the true values of κ and γ by κ^* and γ^*.

Theorem 3. *Under Assumptions 1–3, and denote* $\rho = p(G=1)/\{1-p(G=1)\}$, *which is the limit of* n_1/n_0, *we have*

(i) if the propensity scores e_i*'s in RCT data are known, then*

$$\sqrt{n_1}(\hat{\tau}_{ipw}-\tau) \xrightarrow{d} N\Big(0, V_1 + \rho B_1^T I^{-1}(\kappa^*)B_1\Big), \tag{4.3}$$

where $V_1 = var\{\frac{(T_i-e_i)h_i}{e_i(1-e_i)}|G_i=1\}$, $B_1 = E[\frac{(T_i-e_i)}{e_i(1-e_i)}\cdot h_i'(\kappa^*)|G_i=1]$ *and* $I(\kappa^*)$ *is the Fisher information matrix of* κ *at* κ^* *in observational data.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

(ii) if we estimate the propensity scores e_i's in RCT data with a correctly specified logistic regression model, then

$$\sqrt{n_1}(\tilde{\tau}_{ipw} - \tau) \xrightarrow{d} N\Big(0, (V_1 - V_2) + \rho B_1^T I^{-1}(\kappa^*) B_1\Big), \tag{4.4}$$

where $V_2 = B_2^T I^{-1}(\gamma^*) B_2$, *with* $B_2 = E[\frac{T_i h_i (1-e_i) X_i}{e_i} | G_i = 1] + E[\frac{(1-T_i) h_i e_i X_i}{1-e_i} | G_i = 1]$, $I(\gamma^*) = E[e_i(1-e_i) X_i X_i^T | G_i = 1]$ *is the Fisher information matrix of* γ *at* γ^*.

Theorem 3(i) shows the asymptotic variance of $\sqrt{n_1}\hat{\tau}_{ipw}$ consists of V_1 and $\rho B_1^T I^{-1}(\kappa^*) B_1$, where the former is the variance of IPW estimator when $h(X, S, T)$ is known by noting that $V_1 = \text{var}\{T_i h_i / e(X_i) - (1 - T_i) h_i / (1 - e(X_i)) | G_i = 1\}$, which can be seen as the systematic variance; the latter is induced by the estimation of $h(X_i, S_i, T_i; \kappa)$. Compared with $\hat{\tau}_{ipw}$, the asymptotic variance of $\tilde{\tau}_{ipw}$ in Theorem 3(ii) minus an extra positive term V_2 resulted from the estimation of propensity scores, which reveals that using estimated propensity scores reduces the asymptotic variance and thus lead to a more accurate estimator. This phenomenon has been noticed in previous literature, such as (Joffe and Rosenbaum, 1999; Hirano et al., 2003; Wu et al., 2021), and we will verify it in the simulation study of Section 5.

The results given in Theorem 3 are valid for any generalized linear

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

model $h(X_i, S_i, T_i; \kappa)$, and thus it is applicable to various data types of Y. For convenience, we present the specific form of B_1 and $I(\kappa^*)$ for binary and continuous outcomes, the two most common scenarios in real data analysis. (1) For binary Y and assume $h(X_i, S_i, T_i; \kappa)$ is a logistic model, then $B_1 = E[\frac{(T_i - e_i)}{e_i(1-e_i)} \cdot h_i(1-h_i)\tilde{X}_i | G_i = 1]$, $I(\kappa^*) = E[h_i(1-h_i)\tilde{X}_i\tilde{X}_i^T | G_i = 0]$. (2) For continuous outcome and assume $h(X_i, S_i, T_i; \kappa)$ is a linear model with variance σ^2, then $B_1 = E[\frac{(T_i - e_i)}{e_i(1-e_i)} \cdot \tilde{X}_i | G_i = 1]$, $I(\kappa^*) = E[\tilde{X}_i\tilde{X}_i^T\sigma^{-2} | G_i = 0]$.

Furthermore, with respect to the efficiency between the IPW and the efficient doubly robust estimator, we have the following corollary.

Corollary 1. *When (X, S) are discrete with finite support, and the nuisance parameters in $\hat{\tau}_{dr}$ and $\tilde{\tau}_{ipw}$ are nonparametrically estimated of order $\sqrt{n}$, as the efficient influence function ϕ is the unique influence function for τ, we have $\hat{\tau}_{dr}$ and $\tilde{\tau}_{ipw}$ are first-order equivalent, that is, they have the same asymptotic distribution.*

Corollary 1 shows that the IPW estimator using estimated propensity scores achieves the semiparametric efficient bound in the case of discrete X and S with finite support. The intuition is that when X and S are discrete with finite support, the selection model $g_t(S, X)$, the propensity score model $e(X)$, and the regression model $\mu_t(S, X), \mu_t(X)$ only contain finite dimensional parameters, which can be nonparametrically estimated

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

at a convergence rate of order $1/\sqrt{n}$. Therefore the IPW method with estimated propensity score and the doubly robust estimator is regular and asymptotically normal for the nonparametric model constrained only by Assumptions 1–3. Their corresponding influence function must be the only element ϕ by Theorem 1(ii), so the two estimators have the same asymptotic distribution.

The asymptotic variances of both $\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$ can be obtained by the plug-in method, that is, substitute $e(X_i)$ and h_i with its estimates $\hat{e}(X_i)$ and $\hat{h}_i$ in the associated asymptotic variance formulas, and the population expectation is replaced by the empirical average. In Section 5, the simulation study shows that the estimated asymptotic variances based on the plug-in method perform well across extensive simulation scenarios.

5. Simulation

We conduct extensive simulation studies to assess the finite sample performance of the proposed methods and compare them with the competing approach of Athey et al. (2019). Two common data types of Y, binary and continuous, are considered in this simulation. `R` codes are provided in `https://github.com/hwj0828/long-term-effect` to reproduce the simulation results.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Denote U as the unmeasured variable. Throughout this simulation, for RCT data, unmeasured variable $U \sim N(0,1)$, and the treatment assignment $\text{pr}(T=1)=1/2$. The error terms ε_S and ε_Y are independently identically distributed in $N(0,1)$ for both the RCT and observational data. The sample size of RCT data is set as $n_1 = 50, 100$ or 200, observational data is $n_0 = 500$. Let $\text{expit}(x) = \exp(x)/\{1+\exp(x)\}$ be the logistic function.

Continuous outcome. We first consider the following four cases for continuous Y.

Case (1). For RCT data , $S = U + 2(X_1 + X_2) + T + \varepsilon_S$, $Y = T + 3(X_1 + X_2) + S + \varepsilon_Y$, $X = (X_1, X_2)^T \sim N(0, I_2)$. For observational data, $X \sim N(1, 4I_2), U \sim N(0,1)$, $\text{pr}(T=1|X,U) = \text{expit}\{U + X_1 + X_2\}$, S and Y are generated the same as in RCT data .

Case (2). For RCT data , $S = U^2 + 2(X_1^2 + X_2^2) + T + \varepsilon_S$, $Y = T + 3(X_1 + X_2) + S + \varepsilon_Y$, $X = (X_1, X_2)^T \sim N(0, I_2)$. For observational data, $X \sim N(1, 4I_2), U \sim N(0,1)$, $\text{pr}(T=1|X,U) = \text{expit}\{U + X_1 + X_2U\}$, S and Y are generated the same as in RCT data .

Case (3). The data generation mechanism is the same as in case (1), except for setting $U \sim N(1,4)$ in observational data.

Case (4). The data generation mechanism is the same as in case (2), except for setting $U \sim N(1,4)$ in observational data.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

The unmeasured confounder U influences both T and S in observational data for all cases (1)-(4). The distribution of U in RCT data is the same as that in observational data for cases (1) and (2) and is different from that in observational data for cases (3) and (4). As discussed in Section 2, Assumptions 1-3 hold for all cases (1)-(4), while Assumption 4 holds only for cases (1)-(2) and is violated for cases (3)-(4). In addition, to mimic the real-world data, we set the distributions of covariates between the RCT and observational data to be different for all cases (1)-(4), and the surrogates S may be a linear function of (T, X, U) (cases (1) and (3)), or a non-linear function of (T, X, U) (cases (2) and (4)).

Each simulation study is based on 1000 replicates. In the following tables, Bias and SD are the Monte Carlo bias and standard deviation over the 1000 simulations of the points estimates. ESE and CP95 are the averages of estimated asymptotic standard error and coverage proportions of the 95% confidence intervals based on the plug-in method, respectively. ESE.b and CP95.b have the same meaning as ESE and CP95 but are derived from 200 bootstraps. The true value of τ is obtained by generating RCT data with a sample size of 100000.

Table 1 summarizes the numeric results of the proposed estimators (doubly robust estimator $\hat{\tau}_{dr}$, IPW estimators $\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$) and the com-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Table 1: Comparison of various estimators for cases (1)-(4), continuous outcome. The distributions of U between the RCT and observational data are the same for cases (1)-(2) and different for cases (3)-(4).

	$n_1 = 50$					$n_1 = 100$					$n_1 = 200$				
Case	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b
IPW Estimator ($\hat{\tau}_{ipw}$), with True Propensity Score															
(1)	8.1 (210.5)	206.2	93.3	202.3	92.7	8.7 (144.8)	146.6	94.9	145.4	94.0	-2.1 (105.2)	103.5	93.7	102.5	92.8
(2)	1.7 (240.0)	239.3	94.7	235.9	94.2	-7.4 (176.4)	171.3	93.9	169.3	93.1	2.1 (119.5)	121.2	95.8	121.0	95.8
(3)	13.1 (206.6)	206.6	93.9	202.9	93.6	1.1 (147.5)	145.2	95.3	144.1	95.0	9.8 (100.2)	103.4	95.7	102.2	95.2
(4)	7.0 (250.0)	239.5	93.9	236.3	93.3	13.6 (168.2)	171.4	95.9	169.8	95.1	8.2 (120.1)	121.6	94.9	121.0	94.5
IPW Estimator ($\hat{\tau}_{ipw}$), with Estimated Propensity Score															
(1)	5.0 (51.6)	55.6	96.4	92.1	98.6	2.9 (32.4)	33.9	96.1	39.4	97.1	2.1 (23.9)	24.4	95.0	25.4	95.1
(2)	1.2 (138.8)	125.9	91.4	163.7	95.4	0.7 (94.1)	89.5	94.1	97.4	95.4	2.4 (66.2)	63.5	94.2	65.4	94.7
(3)	8.0 (52.1)	55.6	96.4	95.8	99.4	7.0 (32.3)	34.4	95.3	40.1	97.2	8.4 (23.8)	24.8	94.3	25.8	93.6
(4)	3.2 (150.0)	128.0	92.3	174.9	95.8	7.9 (95.2)	89.5	94.3	97.4	95.7	7.6 (66.0)	63.6	93.9	65.3	94.0
Doubly Robust Estimator ($\hat{\tau}_{dr}$)															
(1)	-1.5 (41.9)	48.7	96.8	39.4	91.8	0.3 (30.9)	35.9	97.4	30.3	94.3	0.8 (23.8)	28.0	97.4	24.2	94.9
(2)	-4.7 (116.7)	118.1	94.8	104.0	91.6	-1.4 (86.3)	87.5	95.9	80.9	92.9	0.9 (63.9)	63.8	95.3	60.7	93.3
(3)	2.3 (41.2)	49.0	98.2	39.5	93.2	4.9 (30.7)	37.1	98.6	31.1	95.4	6.9 (25.3)	30.7	98.4	25.5	93.6
(4)	-0.6 (123.3)	119.3	95.1	104.9	91.4	4.1 (88.1)	88.8	95.6	82.2	93.2	5.9 (65.8)	65.4	95.3	61.9	92.5
Athey et al. (2019)'s Method															
(1)	-83.2 (214.4)	212.1	93.6	206.3	92.0	-83.1 (147.3)	149.9	91.9	148.2	92.0	-93.9 (108.0)	106.4	86.0	106.7	85.4
(2)	-101.2 (242.7)	243.5	92.9	238.6	91.9	-112.1 (178.0)	173.5	88.6	171.2	87.1	-108.9 (121.1)	122.5	84.1	122.2	83.8
(3)	-80.7 (210.1)	211.5	92.1	206.2	92.0	-94.3 (149.6)	148.2	89.3	147.3	88.2	-90.4 (124.9)	114.6	87.1	124.2	86.8
(4)	-99.5 (254.5)	245.3	91.6	240.8	89.7	-99.0 (170.8)	175.0	91.8	173.2	91.0	-115.4 (123.0)	124.9	83.3	124.9	82.9

Note: All the values in this table have been magnified 100 times. Bias and SD are the Monte Carlo bias and standard deviation over the 1000 simulations of the points estimates. ESE and CP95 are the averages of estimated asymptotic standard error and coverage proportions of the 95% confidence intervals based on the plug-in method, respectively. ESE.b and CP95.b have the same meaning as ESE and CP95 but are derived from 200 bootstraps.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

peting estimator in Athey et al. (2019) for cases (1)-(4). For all the proposed estimators $\hat{\tau}_{dr}$, $\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$, the Bias is small, ESE is close to SD and CP95 is close to its nominal value of 0.95. This shows the validity of the asymptotic variance estimation using the plug-in method. As expected, $\hat{\tau}_{dr}$ and $\tilde{\tau}_{ipw}$ have better performance than the $\hat{\tau}_{ipw}$ in terms of smaller Bias and SD. Remarkably, the results of the IPW estimator with estimated propensity score are similar to those of the doubly robust estimator. In addition, the method of Athey et al. (2019) has a significantly larger Bias than the other three approaches, and its CP95 is less than 0.95. A possible reason is that Athey et al. (2019)'s method does not allow T to have a direct effect on Y, whereas in our setups we set T to have a direct effect on Y.

In the setup of simulation study, we assume that the treatment has a direct effect on the outcome. However, , see Assumption 2 and the associated discussion in Athey et al. (2019).

To verify the conclusion of Corollary 1, we conduct two additional simulation scenarios (cases (5)-(6)). The data generating process of cases (5)-(6) are the same as cases (1)-(2), respectively, except that setting X and S as discrete variables. For cases (5)-(6), the covariates X consisting of two binary variables (X_1, X_2), X_1 and X_2 are independent and identically distributed from a Binomial distribution $B(1, 0.5)$ for both the RCT and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Table 2: Comparison of various estimators for cases (5)-(6), continuous outcome, discrete X and S.

	$n_1 = 50$					$n_1 = 100$					$n_1 = 200$				
Case	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b
IPW Estimator ($\hat{\tau}_{ipw}$), with True Propensity Score															
(5)	-1.3 (122.1)	120.4	94.2	118.9	93.2	1.2 (85.1)	85.7	94.6	85.2	94.3	-1.4 (58.5)	61.0	96.0	60.6	95.0
(6)	8.2 (123.0)	123.7	94.4	121.9	93.4	4.1 (88.6)	87.8	93.7	86.8	92.5	-0.9 (63.5)	62.6	94.6	62.1	94.1
IPW Estimator ($\hat{\tau}_{ipw}$), with Estimated Propensity Score															
(5)	-1.5 (16.7)	18.0	96.2	26.0	98.7	-0.7 (12.5)	13.2	95.1	14.0	96.3	-0.7 (11.4)	11.7	95.9	11.7	95.9
(6)	-0.2 (17.5)	18.6	96.3	25.7	98.9	-0.2 (13.3)	13.7	94.9	14.4	95.6	-0.3 (11.6)	11.8	94.9	11.9	94.6
Doubly Robust Estimator ($\hat{\tau}_{dr}$)															
(5)	-2.0 (15.3)	21.3	98.3	15.3	94.7	-1.0 (12.6)	16.3	98.3	13.2	95.7	-0.9 (11.8)	13.1	97.2	12.0	95.5
(6)	-0.8 (16.1)	22.1	98.7	16.0	94.9	-0.6 (13.4)	16.8	98.2	13.5	94.2	-0.4 (11.7)	13.4	97.4	12.1	94.5
Athey et al. (2019)'s Method															
(5)	-97.8 (127.6)	128.5	88.5	124.8	85.7	-94.7 (88.5)	90.3	83.0	89.0	81.5	-97.4 (60.8)	63.5	66.9	62.9	65.8
(6)	-89.8 (128.9)	131.8	89.4	127.6	88.2	-93.7 (92.7)	92.3	81.9	90.6	81.0	-99.4 (65.9)	65.1	67.9	64.3	66.7

observational data. The surrogate S is generated through a logistic regression with $\text{pr}(S = 1|X, S, U, T) = \text{expit}\{U - 2(X_1 + X_2) + T\}$ for case (5) and $\text{pr}(S = 1|X, S, U, T) = \text{expit}\{U^2 - 2(X_1^2 + X_2^2) + T\}$ for case (6). Table 2 shows the simulation results for cases (5)-(6). As expected, the IPW estimator with estimated propensity scores and the doubly robust estimator have similar performance for discrete X and S.

Binary outcome. Corresponding to cases (1)-(6), we set 6 simulations (cases (7)-(12)) to evaluate the performance of the proposed methods with binary outcomes. The data generation mechanisms for cases (7)-(12) are provided in Section 5.1 of Supplementary Material. Tables 3-4 summarize the numeric results for cases (7)-(12).

The results presented in Table 3 are similar to those in Table 1, other than the CP95 of the doubly robust estimator is significantly lower than

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Table 3: Comparison of various estimators for cases (7)-(10), binary outcome. The distributions of U between the RCT and observational data are the same for cases (7)-(8) and different for cases (9)-(10).

	$n_1 = 50$					$n_1 = 100$					$n_1 = 200$				
Case	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b
IPW Estimator ($\hat{\tau}_{ipw}$), with True Propensity Score															
(7)	0.0 (20.7)	20.3	93.7	20.2	92.3	0.0 (14.7)	14.7	95.0	14.7	94.0	-0.4 (10.7)	10.6	94.2	10.7	93.8
(8)	0.1 (25.0)	25.6	95.2	25.3	94.3	0.4 (18.2)	18.3	95.2	18.3	94.6	-0.2 (12.9)	13.3	95.5	13.3	95.0
(9)	0.9 (19.6)	20.5	95.0	20.3	94.4	0.4 (15.0)	14.8	93.9	14.8	93.9	0.0 (10.8)	10.8	94.8	10.9	94.6
(10)	-0.2 (24.5)	27.3	95.9	25.7	94.7	1.6 (18.8)	18.7	94.5	18.7	93.4	0.5 (14.3)	13.8	93.3	13.8	92.9
IPW Estimator ($\hat{\tau}_{ipw}$), with Estimated Propensity Score															
(7)	0.0 (8.0)	7.9	94.3	9.7	96.9	0.0 (6.3)	6.0	94.0	6.4	94.8	-0.1 (5.0)	4.9	93.6	5.2	94.4
(8)	0.1 (9.6)	8.5	95.8	11.2	98.5	0.8 (6.1)	6.1	94.5	6.7	95.9	0.7 (5.2)	5.1	94.4	5.3	94.5
(9)	0.0 (8.2)	8.3	95.1	10.3	97.5	0.4 (6.3)	6.3	94.8	6.7	95.6	0.1 (5.4)	5.3	93.7	5.5	94.7
(10)	0.3 (8.8)	10.4	95.1	12.0	98.1	0.7 (7.4)	6.9	92.4	7.6	94.5	0.5 (6.5)	6.1	92.6	6.5	93.8
Doubly Robust Estimator ($\hat{\tau}_{dr}$)															
(7)	-0.3 (7.8)	7.1	93.1	8.3	95.6	-0.1 (6.5)	5.2	88.4	6.7	95.0	-0.1 (5.3)	3.9	86.4	5.5	94.8
(8)	0.2 (7.5)	7.6	95.0	8.3	95.9	0.4 (6.3)	6.0	94.5	6.9	95.9	0.4 (5.6)	5.0	92.8	5.8	95.1
(9)	0.0 (8.1)	7.1	92.1	8.7	95.2	0.2 (6.5)	5.3	89.1	7.2	95.4	0.0 (5.8)	4.1	82.8	6.0	94.7
(10)	-0.5 (9.3)	8.6	94.5	9.9	96.4	-0.3 (8.3)	7.0	90.8	8.1	94.1	-0.2 (7.2)	5.9	89.3	7.2	93.1
Athey et al. (2019)'s Method															
(7)	-4.4 (20.5)	20.2	94.1	19.8	92.4	-4.7 (14.2)	14.2	93.5	14.2	92.7	-4.9 (10.1)	10.0	91.5	10.0	91.0
(8)	-6.0 (24.3)	25.3	95.1	24.7	93.7	-5.8 (17.4)	17.8	94.2	17.6	93.9	-6.5 (12.4)	12.6	92.2	12.5	91.8
(9)	-4.0 (19.4)	20.3	95.4	19.9	93.9	-4.5 (14.6)	14.2	94.0	14.1	92.8	-4.7 (9.9)	10.0	92.7	10.0	92.0
(10)	-6.1 (23.8)	25.1	95.7	24.8	94.4	-4.6 (18.0)	17.8	93.4	17.8	92.3	-5.8 (13.2)	13.0	91.5	12.7	90.5

Table 4: Comparison of various estimators for cases (11)-(12), binary outcome, discrete X and S.

	$n_1 = 50$					$n_1 = 100$					$n_1 = 200$				
Case	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b	Bias (SD)	ESE	CP95	ESE.b	CP95.b
IPW Estimator ($\hat{\tau}_{ipw}$), with True Propensity Score															
(11)	0.3 (26.8)	26.5	93.5	26.1	92.8	0.7 (18.7)	18.7	94.9	18.5	93.8	0.3 (13.1)	13.3	94.7	13.1	94.6
(12)	-0.3 (25.7)	26.7	95.8	26.3	95.0	0.1 (18.7)	18.9	94.9	18.7	94.2	-0.2 (14.2)	13.4	93.2	13.3	93.2
IPW Estimator ($\hat{\tau}_{ipw}$), with Estimated Propensity Score															
(11)	0.0 (3.4)	3.7	97.0	5.7	99.4	0.0 (2.5)	2.6	95.0	2.8	96.2	0.1 (2.1)	2.1	96.0	2.1	95.8
(12)	0.1 (3.5)	3.5	97.0	5.4	99.7	0.0 (2.4)	2.4	95.5	2.6	96.4	0.1 (1.9)	2.0	95.4	2.0	95.4
Doubly Robust Estimator ($\hat{\tau}_{dr}$)															
(11)	-0.1 (3.2)	3.3	94.6	3.2	93.1	0.0 (2.5)	2.7	95.7	2.6	95.4	0.0 (2.2)	2.3	95.9	2.2	94.5
(12)	0.1 (2.9)	3.1	95.6	2.9	93.7	0.0 (2.4)	2.5	95.1	2.4	93.2	0.1 (2.0)	2.2	96.8	2.1	94.8
Athey et al. (2019)'s Method															
(11)	-4.5 (26.6)	27.3	94.8	26.0	93.0	-4.2 (18.5)	19.0	94.9	18.4	93.8	-4.6 (13.0)	13.3	94.4	13.1	93.9
(12)	-4.9 (25.8)	27.4	96.0	26.3	94.9	-4.5 (18.7)	19.1	94.6	18.7	92.9	-4.8 (14.1)	13.4	92.3	13.2	91.2

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

0.95 for all cases (7)-(10). In comparison, the CP95 of IPW estimators $\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$ still have good performance. This indicates the asymptotic variance estimation of IPW estimators based on the plug-in method is more robust than that of the doubly robust estimator. A possible reason is that the estimation of the doubly robust estimator relies on many parametric model specifications and the asymptotic variance formula based on the plug-in method is valid only when all the models are correctly specified. Nonetheless, the CP95.b of both the doubly robust estimator and IPW estimators performs well in all cases (1)-(12), which means that bootstrap can produce more robust variance estimates than the plug-in method, at a high computational cost. Table 4 shows the results for discrete X and S, again demonstrating the equivalence of the IPW estimator with estimated propensity score and the doubly robust estimator.

We further explore the finite sample behaviors of the proposed estimators in the scenarios where the unmeasured confounder U affects both S and Y in observational data. In this case, Assumption 3 may not be satisfied. The corresponding numeric results are similar to those in Tables 1 and 3 and are presented in Tables S1 and S2 of Supplementary Material.

In summary, the simulation results reveal the following phenomena: (1) the IPW estimators ($\hat{\tau}_{ipw}$ and $\tilde{\tau}_{ipw}$) have more stable performance than

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

the other two estimators, a possible reason is that IPW estimators rely on parsimonious model specifications; (2) using the estimated propensity scores can significantly improve the efficiency of IPW estimator; (3) the doubly robust estimator has similar performance to IPW estimator with estimated propensity score concerning Bias and SD. (4) the method of Athey et al. (2019) is less attractive in terms of both Bias and SD.

6. Real data analysis

Immunoglobulin A nephropathy (IgAN), also called Berger disease, is the most prevalent chronic and primary glomerular disease worldwide (Haas, 1997; D'amico, 1987). The renin-angiotensin-aldosterone system (RAAS) inhibition is a standard therapy for IgAN disease by slowing proteinuria and lowing blood pressure (Zhang, Lu, Feng, Li, and Wang, Zhang et al.). Despite the usage of RAAS, lgAN patients are still at risk of renal failure (Liu et al., 2019). Hydroxychloroquine (HCQ), an immunomodulator, is a current therapeutic option for lgAN. Evidence suggests that combination therapy with HCQ and RAAS is effective in reducing proteinuria in patients with IgAN compared to RAAS alone over 6 months (Yang et al., 2018). However, the long-term effect of HCQ on renal outcomes is less clear (Zhang, Lu, Feng, Li, and Wang, Zhang et al.). This study aims to explore the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

treatment effect of HCQ on renal failure by combing an RCT dataset and an observational dataset obtained from Peking University First Hospital.

The RCT data come from a double-blind, randomized, and placebo-controlled trial consisting of 60 observations, of which 30 patients are assigned to the combination therapy with HCQ and RAAS and the rest are assigned to standard RAAS therapy. More details of the RCT data can be found in Liu et al. (2019). The observational data contain 547 observations, of which 91 patients accept the combination therapy. The endpoint (outcome) of interest is a binary variable indicating whether a patient developed renal failure within 3, 4, or 5 years. In this analysis, we consider two endpoints. `endpoint 1` is defined as whether glomerular filtration rate (GFR) decreased by 30%, 40%, or 50% from baseline to the end time, `endpoint 2` is an indicator of whether the GFR is less than 15 ml/min per 1.73 m^2. Since the randomized controlled trial lasted only six months, no endpoints were observed in RCT data. The surrogate is chosen as the percentage change in proteinuria between baseline and six months. The baseline covariates are the same between the RCT data and observational data, including gender, age, baseline proteinuria, baseline GFR, and some pathologic predictors of renal failure (Shi et al., 2011).

The analysis in Liu et al. (2019) shows that the new therapy has better

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Table 5: Estimated effects of HCQ on renal failure.

End time		proportion = 0.3		proportion = 0.4		proportion = 0.5	
		Endpoint 1	Endpoint 2	Endpoint 1	Endpoint 2	Endpoint 1	Endpoint 2
		IPW Estimator ($\hat{\tau}_{ipw}$), with True Propensity Score					
3	Estimate (ESE.b)	-0.376 (0.11)	-0.202 (0.098)	-0.162 (0.065)	-0.202 (0.103)	-0.123 (0.071)	-0.202 (0.100)
	p-value	$< 10^{-3}$	0.020	0.006	0.024	0.041	0.022
4	Estimate (ESE.b)	-0.418 (0.105)	-0.157 (0.085)	-0.211 (0.073)	-0.157 (0.084)	-0.172 (0.083)	-0.157 (0.077)
	p-value	$< 10^{-3}$	0.033	0.002	0.032	0.019	0.021
5	Estimate (ESE.b)	-0.488 (0.109)	-0.192 (0.080)	-0.265 (0.083)	-0.192 (0.080)	-0.211 (0.072)	-0.192 (0.084)
	p-value	$< 10^{-3}$	0.008	$< 10^{-3}$	0.008	0.002	0.011
		IPW Estimator ($\tilde{\tau}_{ipw}$), with Estimated Propensity Score					
3	Estimate (ESE.b)	-0.318 (0.082)	-0.178 (0.087)	-0.138 (0.058)	-0.178 (0.089)	-0.102 (0.066)	-0.178 (0.092)
	p-value	$< 10^{-3}$	0.020	0.009	0.023	0.062	0.027
4	Estimate (ESE.b)	-0.36 (0.079)	-0.145 (0.082)	-0.182 (0.063)	-0.145 (0.089)	-0.155 (0.063)	-0.145 (0.075)
	p-value	$< 10^{-3}$	0.040	0.002	0.051	0.007	0.026
5	Estimate (ESE.b)	-0.423 (0.082)	-0.178 (0.082)	-0.232 (0.062)	-0.178 (0.074)	-0.189 (0.065)	-0.178 (0.069)
	p-value	$< 10^{-3}$	0.015	$< 10^{-3}$	0.008	0.002	0.005
		Doubly Robust Estimator ($\hat{\tau}_{dr}$)					
3	Estimate (ESE.b)	-0.256 (0.108)	-0.187 (0.069)	-0.073 (0.102)	-0.187 (0.076)	-0.034 (0.097)	-0.187 (0.073)
	p-value	0.009	0.004	0.237	0.007	0.362	0.005
4	Estimate (ESE.b)	-0.337 (0.090)	-0.148 (0.065)	-0.125 (0.112)	-0.148 (0.060)	-0.092 (0.099)	-0.148 (0.062)
	p-value	$< 10^{-3}$	0.011	0.132	0.006	0.176	0.008
5	Estimate (ESE.b)	-0.396 (0.088)	-0.195 (0.062)	-0.169 (0.116)	-0.195 (0.065)	-0.119 (0.100)	-0.195 (0.064)
	p-value	$< 10^{-3}$	$< 10^{-3}$	0.074	0.001	0.118	0.001
		Athey et al. (2019)'s Method					
3	Estimate (ESE.b)	-0.034 (0.135)	-0.05 (0.084)	-0.045 (0.132)	-0.05 (0.093)	-0.03 (0.082)	-0.05 (0.092)
	p-value	0.400	0.276	0.367	0.296	0.355	0.293
4	Estimate (ESE.b)	-0.039 (0.124)	-0.043 (0.071)	-0.071 (0.177)	-0.043 (0.086)	-0.046 (0.100)	-0.043 (0.085)
	p-value	0.376	0.274	0.344	0.309	0.322	0.306
5	Estimate (ESE.b)	-0.052 (0.156)	-0.023 (0.090)	-0.082 (0.15)	-0.023 (0.105)	-0.060 (0.096)	-0.023 (0.085)
	p-value	0.368	0.397	0.292	0.412	0.266	0.392

Note: ESE.b is estimated asymptotic standard error based on 200 bootstraps. The p-values are obtained by two-sided test, that is $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

efficacy for the surrogate. In our analysis, we are interested in estimating the average treatment effect for the long-term outcome in RCT data, thus can determine whether the new therapy has better efficacy than the standard therapy. The point estimate and corresponding confidence interval is given in Table 5. We also test the null hypothesis $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$. As Y is a binary outcome, the regression model $\mu_t(X, S)$ is fitted with logistic regression, and $\mu_t(X)$ is fitted by linear regression of $\mu_t(X, S)$ on X. The asymptotic standard errors are obtained based on 200 bootstraps. As shown in Table 5, IPW method and the doubly robust method have similar point estimates while having different standard errors. The point estimates of τ are smaller than zero, indicating the potential benefit of the new therapy against the standard therapy. The p-value for the hypothesis test calculated by IPW and doubly robust methods are smaller than 0.05 in most cases, which means that we can reject the null hypothesis at a significance level of 0.05. Besides, with the end time increasing from 3 years to 5 years, the absolute value of the point estimate of τ becomes bigger, which indicates that the efficacy of the new therapy against the standard therapy amplifies over time. This result shows a similar pattern that is observed in Liu et al. (2019), where the treatment effects on surrogates are analyzed. We also report the results where the asymptotic standard errors are computed with

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

the plug-in method, which are similar to those in Table 5 and are presented in Table S3 of Supplementary Material.

7. Discussion

This article investigates the average causal effects on the long-term outcome. Under weaker assumptions than the existing methods, we derive the semiparametric efficiency bound, propose two new estimators and establish their large sample properties. Both simulation studies and real data analysis demonstrate the advantages of the proposed method compared with competing ones. The proposed approach is suitable for various data types of X, S, and Y and thus has wide application scenarios.

We illustrate the proposed estimators by using generalized linear models to estimate the nuisance parameters. It would be interesting to explore the theoretical properties of the proposed estimators when the nuisance parameters are estimated with machine learning methods (Chernozhukov et al., 2018; Wager and Athey, 2018). When X is high-dimensional, one possible extension is to consider how to obtain valid confidence intervals of the proposed estimators when either the propensity score model or the outcome model is correctly specified (Vermeulen and Vansteelandt, 2015; Tan, 2020; Sun and Tan, 2021; Ning et al., 2020; Wu et al., 2024).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Supplementary Material

Supplementary Material available online includes technical proofs and additional numerical results from the simulation and application.

Acknowledgments

The authors thank the associate editor and anonymous reviewers for their helpful comments and valuable suggestions. This research was supported by the National Natural Science Foundation of China (No. 12301370).

References

Athey, S., R. Chetty, and G. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S., R. Chetty, G. Imbens, and H. Kang (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working Paper 26463, National Bureau of Economic Research.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*(27), 7345–7352.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical studies 147*(1), 59–70.

Chen, H., Z. Geng, and J. Jia (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(5), 919–932.

Chen, J. and D. M. Ritzwoller (2023). Semiparametric estimation of long-term treatment effects. *Journal of Econometrics 237*, 105545.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*, 1–68.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

REFERENCES

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, Varoquaux, and et al. (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science 39*(1), 165–191.

Dahabreh, I. J., S. J. A. Haneuse, J. M. Robins, S. E. Robertson, A. L. Buchanan, E. A. Stuart, and et al. (2021). Study designs for extending causal inferences from a randomized trial to a target population. *American journal of epidemiology 190*(8), 1632–1642.

D'amico, G. (1987). The commonest glomerulonephritis in the world: Iga nephropathy. *QJM: An International Journal of Medicine 64*(3), 709–727.

Ding, S., P. Wu, F. Feng, X. He, Y. Wang, Y. Liao, and et al. (2022). Addressing unmeasured confounder for recommendation with sensitivity analysis. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 305–315.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Gamerman, V., T. Cai, and A. Elsäßer (2019). Pragmatic randomized clinical trials: best practices and statistical guidance. *Health Services and Outcomes Research Methodology 19*(1), 23–35.

Ghassami, A., I. Shpitser, and E. T. Tchetgen (2022). Combining experimental and observational data for identification of long-term causal effects. *arXiv preprint arXiv:2201.10743*.

Haas, M. (1997). Histologic subclassification of iga nephropathy: a clinicopathologic study of 244 cases. *American Journal of Kidney Diseases 29*(6), 829–842.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–331.

Hernán, M. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hünermund, P. and E. Bareinboim (2024). Causal inference and data fusion in econometrics. *The Econometrics Journal*, To Appear.

Imbens, G., N. Kallus, X. Mao, and Y. Wang (2022). Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234*.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference For Statistics Social and Biomedical Science*. Cambridge University Press.

Joffe, M. M. and P. R. Rosenbaum (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology 150*(4), 327–333.

Ju, C. and Z. Geng (2010). Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society: Series B 72*(1), 129–142.

Kallus, N. and X. Mao (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv:2003.12408*.

Kallus, N., A. M. Puli, and U. Shalit (2018). Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 10911–10920.

Kallus, N. and A. Zhou (2018). Confounding-robust policy improvement. In *Advances in neural information processing systems*, pp. 9289–9299.

Lauritzen, S. L., O. O. Aalen, D. B. Rubin, and E. Arjas (2004). Discussion on causality [with reply]. *Scandinavian Journal of Statistics 31*(2), 189–201.

Li, X., W. Miao, F. Lu, and X.-H. Zhou (2021). Improving efficiency of inference in clinical trials with external control data. *Biometrics 79*(1), 394–403.

Liu, L.-J., Y.-z. Yang, S.-F. Shi, Y.-F. Bao, C. Yang, S.-N. Zhu, and et al. (2019). Effects of hydroxychloroquine on proteinuria in iga nephropathy: a randomized controlled trial. *American Journal of Kidney Diseases 74*(1), 15–22.

Miao, W., W. Li, W. Hu, R. Wang, and Z. Geng (2022). Invited commentary: estimation and bounds under data fusion. *American Journal of Epidemiology 191*(4), 674–678.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2015). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Neyman, J. S. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*, 465–472.

Ning, Y., S. Peng, and K. Imai (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika 107*, 533–554.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Statistics in medicine 8(4), 431–440.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology 66*, 688–701.

Shi, S.-F., S.-X. Wang, L. Jiang, J.-C. Lv, L.-J. Liu, Y.-Q. Chen, and et al. (2011). Pathologic predictors of renal outcome and therapeutic efficacy in iga nephropathy: Validation of the oxford classification. *Clinical Journal of the American Society of Nephrology 6*(9), 2175–2184.

Sun, B. and Z. Tan (2021). High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business and Economic Statistics*, (To Appear).

Tan, Z. (2007). Comment: understanding or, ps and dr. *Statistical Science 22*, 560–568.

Tan, Z. (2010). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *The Canadian Journal of Statistics 38*, 609–632.

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

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.

Vermeulen, K. and S. Vansteelandt (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association 110*(511), 1024–1036.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wu, P., S. Han, X. Tong, and R. Li (2024). Propensity score regression for causal inference with treatment heterogeneity. *Statistica Sinica 34*, 747–769.

Wu, P., H. Li, Y. Deng, W. Hu, Q. Dai, Z. Dong, J. Sun, R. Zhang, and X.-H. Zhou (2022). On the opportunity of causal learning in recommendation systems: Foundation, estimation,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

prediction and challenges. In *International Joint Conference on Artificial Intelligence*, pp. 5646–5643.

Wu, P., S. Luo, and Z. Geng (2023). On the comparative analysis of average treatment effects estimation via data combination. *arXiv preprint arXiv:2311.00528*.

Wu, P., Z. Tan, W. Hu, and X.-H. Zhou (2024). Model-assisted inference for covariate-specific treatment effects with high-dimensional data. *Statistica Sinica 34*, 459–479.

Wu, P., X. Xu, X. Tong, Q. Jiang, and B. Lu (2021). Semiparametric estimation for average causal effects using propensity score-based spline. *Journal of Statistical Planning and Inference 212*, 153–168.

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*, 1540–1554.

Yang, S., D. Zeng, and X. Wang. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (3), 575–596.

Yang, S., D. Zeng, and X. Wang (2022). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*.

Yang, Y.-Z., L.-J. Liu, S.-F. Shi, Y.-Q. Chen, J.-C. Lv, and H. Zhang (2018). Effects of hydroxychloroquine on proteinuria in immunoglobulin a nephropathy. *American Journal of Nephrology 47*(3), 145–152.

Zhang, J., X. Lu, J. Feng, H. Li, and S. Wang. *BioMed Research International*, 9171715.

Wenjie Hu

Address: Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

E-mail: huwenjie@pku.edu.cn

Xiao-Hua Zhou

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0006

Address: Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

E-mail: azhou@math.pku.edu.cn

Peng Wu

Address: School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, China

E-mail: pengwu@btbu.edu.cn