Bernoulli **31**(4), 2025, 2987–3012
https://doi.org/10.3150/24-BEJ1835

Data fusion methods for the heterogeneity of treatment effect and confounding function

SHU YANG[1,a], SIYI LIU[1,b], DONGLIN ZENG[2,c] and XIAOFEI WANG[3,d]

[1]*Department of Statistics, North Carolina State University, Raleigh, NC, 27607, U.S.A.,* [a]*syang24@ncsu.edu,* [b]*sliu48@ncsu.edu*
[2]*Department of Biostatistics, University of North Carolina, Chapel Hill, 27599, U.S.A.,* [c]*dzeng@email.unc.edu*
[3]*Department of Biostatistics and Bioinformatics, Duke University, Durham, 27708, U.S.A.,* [d]*xiaofei.wang@duke.edu*

The heterogeneity of treatment effect (HTE) lies at the heart of precision medicine. Randomized controlled trials are gold-standard for treatment effect estimation but are typically underpowered for heterogeneous effects. In contrast, large observational studies have high predictive power but are often confounded due to the lack of randomization of treatment. We show that the observational study, even subject to hidden confounding, may be used to empower trials in estimating the HTE using the notion of confounding function. The confounding function summarizes the impact of unmeasured confounders on the difference between the observed treatment effect and the causal treatment effect, given the observed covariates, which is unidentifiable based only on the observational study. Coupling the trial and observational study, we show that the HTE and confounding function are identifiable. We then derive the semiparametric efficient scores and the integrative estimators of the HTE and confounding function. We clarify the conditions under which the integrative estimator of the HTE is strictly more efficient than the trial estimator. Finally, we illustrate the integrative estimators via simulation and an application.

Keywords: Estimating equation; goodness of fit; over-identification test; semiparametric efficiency; structural model

1. Introduction

Randomized controlled trials are the cornerstone of evidence-based medicine for treatment effect evaluation because randomization of treatment ensures that treatment groups are comparable and biases are minimized. Recently, considerable interest has been in understanding the heterogeneity of treatment effects, a critical path toward personalized medicine (Collins and Varmus, 2015). However, due to eligibility criteria for recruiting patients, the trial sample is often limited in the patient diversity, which renders the trial underpowered to estimate the heterogeneity of treatment effect. On the other hand, large observational studies are increasingly available for research purposes, such as electronic health records, claims databases, and disease registries, with much broader demographic and diversity than trial cohorts. However, they also present challenges such as confounding due to the lack of randomization.

Existing approaches to harmonize evidence from trial and observational studies include meta-analysis (Verde and Ohmann, 2015) and joint analysis of the pooled data (Prentice et al., 2008). As related, Chen, Hong and Tarozzi (2008) proposed an efficient generalized-method-of-moments estimator combining primary and auxiliary samples under missingness at random, i.e., no unmeasured confounding in our context. However, these approaches assume no hidden confounders, which is unlikely to be true in practice. The no unmeasured confounding assumption requires researchers to measure all relevant predictors of treatment and outcome. However, it is always possible that certain important confounders are unavailable in uncontrolled, real-world settings. For example, doctors use patients'

symptoms not captured in the medical charts to assign treatments. Alternatively, certain prognostic factors are measured with errors due to technological limitations. Unmeasured confounding presents a major threat to causal inference from observational studies. Classical approaches to mitigating bias due to unmeasured confounding include instrumental variables (Angrist, Imbens and Rubin, 1996), negative controls (Kuroki and Pearl, 2014), and sensitivity analysis (Robins, Rotnitzky and Scharfstein, 2000). In particular, sensitivity analysis is often recommended to assess the robustness of the study conclusion to no unmeasured confounding. Many authors have implemented sensitivity analysis using the so-called confounding function (Robins, Rotnitzky and Scharfstein, 2000, Yang and Lok, 2018, Kasza, Wolfe and Schuster, 2017); namely, the difference of the potential outcome means between the treatment groups given the measured covariates due to the unmeasured confounders. Because the observational studies carry no information about confounding biases due to unmeasured confounders, the confounding function is not identifiable based solely on the observational studies.

In this paper, we leverage observational studies to improve trial analysis of the heterogeneity of treatment effect (HTE) with a vector of known effect modifiers. We focus on the setting where the transportability of the heterogeneous treatment effect holds from the trial to the observational study but the observational study may be subject to hidden confounding. Transportability is a minimal requirement for data integration and has been considered in a vast literature (e.g., Stuart et al., 2011, Tipton, 2013, Buchanan et al., 2018, Dahabreh et al., 2019). It holds if the sample is randomly selected from the population or treatment effect modifiers are fully captured. We also introduce a new and natural confounding function to capture the impact of unmeasured confounding in observational studies on the difference between the observed treatment effect and the causal treatment effect given measured covariates. Under structural model assumptions, we show that the trial can be leveraged to identify the HTE and confounding function, in contrast to sensitivity analysis.

The identification results motivate a broad class of consistent estimators of the model parameters. However, naive choices lead to inefficient estimators. We derive the semiparametric efficient score combining the two data sources to guide constructing efficient estimators and accelerate the full potential of trial and observational studies. The theoretical task is challenging because of restrictions on the parameters of interest induced from the identification assumptions, such that the existing semiparametric efficiency theory for data integration (e.g., Chen, Hong and Tarozzi, 2008) cannot apply. To overcome the challenges, we translate the restrictions into the likelihood function by re-parameterization and follow the geometric approach (e.g., Bickel et al., 1993, Tsiatis, 2006) to derive the efficient score. Built upon the efficient score, we propose an integrative estimator, which enables a fast root-N rate of convergence under weaker conditions on nuisance function approximation, e.g., using consistent but flexible semiparametric and nonparametric methods. We clarify the conditions under which the gain of efficiency is strictly positive by data integration over the trial-only estimator. The improvement in efficiency arises when certain predictors in the heterogeneous treatment effect function are absent in the confounding function. The formulation of the heterogeneous treatment effect and confounding functions can be grounded in domain expertise or determined through variable selection based on training data. Additionally, we propose goodness-of-fit tests for assessing the structural assumptions based on over-identification tests. A simulation study shows that the integrative estimator outperforms the trial-only estimator in two settings with and without unmeasured confounding in the observational study. In addition, we apply the proposed method to estimate the heterogeneous treatment effect of chemotherapy for non-small cell lung cancer. The proofs of the semiparametric efficient score and its asymptotic properties are presented in the last section of the main paper, with more technical proofs given in the supplementary material (Yang et al., 2025).

Our work is motivated by related works and addresses challenges in the areas of data fusion, projection, and nonparametric structural models, each of which is elaborated as follows.

Data fusion. Yang and Ding (2020) developed integrative causal analyses of the average treatment effect by calibrating auxiliary information from the validation sample to the big main sample with unmeasured confounders for efficiency gain. However, their approach requires the validation and main samples to be comparable in providing consistent estimators of auxiliary parameters. This requirement may be stringent because randomized controlled trials often have strict inclusion and exclusion criteria lending their patient compositions different from the observational population (Stuart, Bradshaw and Leaf, 2015). Yang et al. (2023) pretested the comparability between trial and observational studies and customized the subsequent analysis based the pretest result. Another line of research for combining trial and observational studies is to generalize the average treatment effect from trials to the target population (Buchanan et al., 2018, Lee et al., 2023; Lee, Yang and Wang, 2022), where the observational sample provides a representative covariate distribution of the target population. Yang and Wang (2022) and Colnet et al. (2024) provided comprehensive reviews. As a by-product of the proposed framework, we derive an efficient plug-in sample estimator of the population average treatment effect in the data integration context. Most existing methods rely on the overlap assumption of the covariate distribution between the trial and observational samples. Our method does not require the overlap assumption but utilizes parametric structural assumptions on the HTE, thus offering an alternative means for causal generalization. Nonetheless, we add a caveat that the lack of overlap renders the structural assumptions fragile, and one relies on model extrapolation. In practice, we still advocate checking the overlap assumption for generalizing the treatment effect from trial to a target population.

Projection parameters. In clinical applications, parametric models are preferable for their straightforward interpretability. We operate under the assumption that the structural models are parametric and are correctly specified. However, in cases of misspecification, they can still be interpreted as projection parameters—the projection of nonparametric structural models onto a constrained model space. Projection-based interpretation has gained popularity and facilitates deriving nonparametric efficient scores for the projection parameters (e.g., Neugebauer and van der Laan, 2005, Chernozhukov et al., 2018a, Kennedy, Lorch and Small, 2019, Kennedy, Balakrishnan and Wasserman, 2023). Unlike these approaches, our semiparametric efficiency scores take into account the constraints on the structural parameters imposed by the identification assumptions. Semiparametric efficiency for models subject to constraints can be of independent interest.

Nonparametric structural models. Aside from the interpretability, a technical reason to consider parametric models is that the nonparametric HTE and confounding functions are often local parameters that are not pathwise differentiable, so their efficient scores with finite variances do not exist (Bickel et al., 1993). Exceptions include the cases of fully discrete data and functions valued in general Hilbert space (Luedtke and Chung, 2024). With the nonparametric models, one can alternatively study the bounds on the asymptotic minimax risk indicating the best possible performance of any estimator in the worst case scenarios (Kennedy et al., 2024 and references therein). For example, Kennedy et al. (2024) derived a lower bound on the minimax rate of HTE estimation when the HTE and nuisance functions are Holder-smooth in studies without hidden confounding. When considering nonparametric rates of HTE estimation in the data fusion context, one strategy involves initially approximating the HTE by projecting it onto a finite-dimensional approximating space and then balancing estimation accuracy and approximation error. However, in such analyses, the objective differs from what is presented in this paper; specifically, there is no need to derive a precise asymptotic distribution of the estimation for the finite-dimensional projection. Instead, the focus is on controlling the mean-squared error as the dimension of the projection increases. This research topic will be pursued in the future.

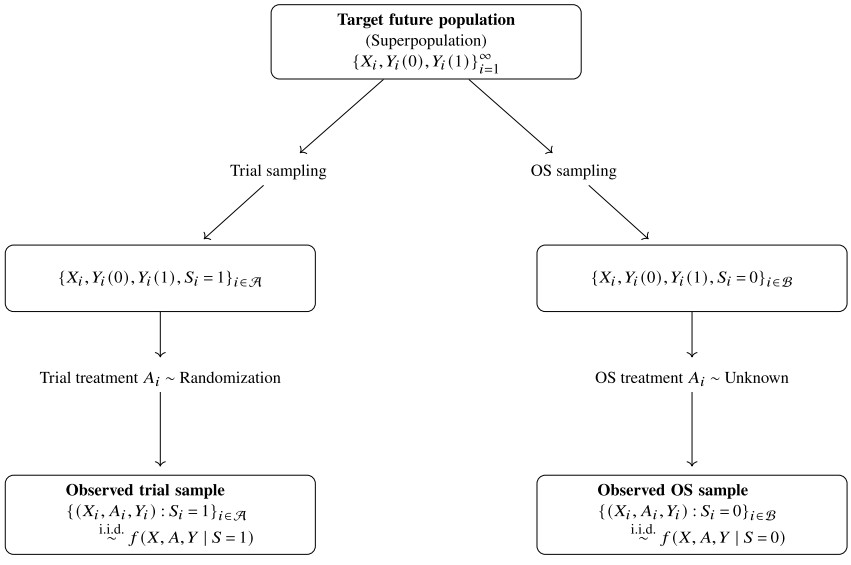

Figure 1. Demonstration of the data structure for the trial and observational study (OS) samples within the target population.

2. Basic setup and identification

2.1. Notation, causal effects, and two data sources

Let A be the binary treatment, X be a vector of pre-treatment covariates with the first component being 1, and Y be the outcome of interest. The target population consists of all patients with certain diseases where the new treatment is intended to be given. We use the potential outcomes to define causal effects. Let $Y(a)$ be the potential outcome had the subject been given treatment a, for $a = 0, 1$. Based on the potential outcomes, the individual treatment effect becomes $Y(1) - Y(0)$, the HTE can be characterized through $\tau(X) = \mathrm{E}[Y(1) - Y(0) \mid X]$, and the average treatment effect is $\tau_0 = \mathrm{E}[\tau(X)]$.

We consider two independent data sources: one is a randomized trial study, and the other is an observational study. Let $S = 1$ denote trial participation, and let $S = 0$ denote observational study participation. Let $\mathcal{A}$ and $\mathcal{B}$ be sample index sets for the two data sources with sample sizes $|\mathcal{A}| = n$ and $|\mathcal{B}| = m$, and the total sample size is $N = n + m$. The trial data consist of $\{V_i = (A_i, X_i, Y_i, S_i) : i \in \mathcal{A}, S_i = 1\}$, where the observations i.i.d. follow $f(X, A, Y \mid S = 1)$, and the observational data consist of $\{V_i : i \in \mathcal{B}, S_i = 0\}$, where the observations i.i.d. follow $f(X, A, Y \mid S = 0)$. Figure 1 displays the envisioned data structure within the target population. To link the observed outcome and potential outcomes, we make the typical causal consistency assumption of $Y = Y(A)$. This assumption rules out the interference between subjects and treatment version relevance between samples and population (Tipton, 2013). The implication is that $Y(a)$ has consistent meaning and value across the trial and observational studies. This assumption requires the same treatment or comparison conditions to be given to both studies, and being in the trial should not affect the values of the potential outcomes. To simplify the exposition, we define

$$e(X, S) = \mathrm{P}(A = 1 \mid X, S), \mu_a(X, S) = \mathrm{E}[Y \mid A = a, X, S],$$

$$\sigma_a^2(X, S) = \mathrm{var}[Y \mid A = a, X, S], \mu(X, S) = \mathrm{E}[Y \mid X, S],$$

where $e(X,S)$ is the propensity score, $\mu_a(X,S)$ and $\sigma_a^2(X,S)$ are the treatment-specific outcome mean and variance functions, for $a=0,1$, and $\mu(X,S)$ is the outcome mean function marginalized over treatment. For any $g(V)$, define $\epsilon_{g(V)} = g(V) - \mathrm{E}[g(V) \mid X,S]$, e.g., $\epsilon_A = A - e(X,S)$.

2.2. Assumptions, confounding function, and nonparametric identification

Due to the fundamental problem that the potential outcomes can never be jointly observed for a particular subject, $\tau(X)$ is not identifiable in general. We make the following assumptions.

Assumption 1 (Transportability and randomized trial design). (i) $\mathrm{E}[Y(1) - Y(0) \mid X, S=s] = \tau(X)$, for $s=0,1$, (ii) $Y(a) \perp\!\!\!\perp A \mid (X, S=1)$ for $a \in \{0,1\}$, and $0 < e(X,S) < 1$ almost surely.

Assumption 1(i) states that the treatment effect function is transportable from the trial and observational samples to the target population. This assumption is common in the data integration literature. Stronger versions of Assumption 1(i) include the ignorability of study participation (e.g., Buchanan et al., 2018) and the mean exchangeability (e.g., Dahabreh et al., 2019). Assumption 1(i) holds if X captures the heterogeneity of effect modifiers or the study sample is a random sample from the target population. To ensure this assumption holds, variables and samples should be carefully chosen with consultations of subject knowledge; e.g., collect data on likely effect modifiers that affect study participation. Under the structural equation model framework, Pearl and Bareinboim (2011) provided graphical conditions for transportability. Assumption 1(ii) holds by a well-designed trial with good patient compliance.

Unlike trials, treatment randomization is typically unrealistic for observational studies. To take into account the possible unmeasured confounders, we define the confounding function

$$\lambda(X) = \mu_1(X, S=0) - \mu_0(X, S=0) - \tau(X),$$

which measures the difference between the observed treatment effect and the causal treatment effect given X. In the absence of unmeasured confounders, we have $\lambda(X) = 0$. In the presence of unmeasured U that is related to both $\{Y(0), Y(1)\}$ and A after controlling for X, we have $\lambda(X) \neq 0$.

Proposition 1. *Under Assumption 1, $\tau(X)$ and $\lambda(X)$ are identifiable by*

$$\tau(X) = \mu_1(X, S=1) - \mu_0(X, S=1) = \mathrm{E}[\widetilde{Y} \mid X, S=1], \tag{1}$$

$$\lambda(X) = \mu_1(X, S=0) - \mu_0(X, S=0) - \tau(X) = \mathrm{E}[\widetilde{Y} \mid X, S=0] - \tau(X), \tag{2}$$

where $\widetilde{Y} = Y\{A - e(X,S)\}/[e(X,S)\{1 - e(X,S)\}]$.

That is, the transportability and randomized trial design identify $\tau(X)$. The observed treatment effect from the observational study is attributable to both $\tau(X)$ and $\lambda(X)$, but coupling the trial and observational samples identifies $\lambda(X)$. The second equality on (1) and (2) follows by $\mathrm{E}(\widetilde{Y} \mid X, S) = \mu_1(X,S) - \mu_0(X,S)$. Proposition 1 provides two identification strategies relying on different components of the observed data distribution, one using the outcome mean functions and the other using the propensity score via $\widetilde{Y}$.

2.3. Parametric structural models and identification results

In clinical settings, the parametric models of the HTE are desirable due to their easy interpretation. These models offer a transparent way of describing how the treatment effect varies across patients' characteristics and can be used to tailor the treatment to an individual's characteristics (Chakraborty and Moodie, 2013). We make the following parametric structural assumptions.

Assumption 2 (Parametric structural models). The HTE and confounding functions are

$$\tau(X) = \tau_{\varphi_0}(X), \quad \lambda(X) = \lambda_{\phi_0}(X), \tag{3}$$

where $\tau_\varphi(X)$ and $\lambda_\phi(X)$ are known continuous functions of $\varphi \in \Theta_1$ and $\phi \in \Theta_2$, φ_0 and ϕ_0 are unique but unknown values, and Θ_1 and Θ_2 are compact sets in $\mathcal{R}^{p_1}$ and $\mathcal{R}^{p_2}$, respectively.

When X is discrete and models are saturate, the structural models are nonparametric. The continuity of $\tau_\varphi(X)$ and $\lambda_\phi(X)$ and compactness of Θ_1 and Θ_2 are imposed for identification and are standard in the literature. The treatment effect model $\tau_{\varphi_0}(X)$ is a special case of structural nested mean models (Robins, 1994) with a single treatment. Tian et al. (2014) and Vansteelandt and Joffe (2014) considered a linear treatment effect $\tau_{\varphi_0}(X) = X^{\mathrm{T}}\varphi_0$, where the first component of X is one, specifying an intercept term. This model entails that on average, the treatment would increase the mean of the outcome by $X^{\mathrm{T}}\varphi_0$, and the magnitude of the increase depends on X. Moreover, each component of φ_0 quantifies the magnitude of the treatment effect of each component of X. Assume that higher values are indicative of better outcomes. If $X^{\mathrm{T}}\varphi_0 > 0$, it indicates that the treatment is beneficial for the subject with X. Other flexible models can also be considered, such as single-index models (Song et al., 2017) and multiple-index models (Chen, Hall and Müller, 2011). Modeling $\lambda_{\phi_0}(X)$ follows the large sensitivity analysis literature (Robins, Rotnitzky and Scharfstein, 2000), which typically requires domain knowledge to identify the possible unmeasured confounders and their relationships with the observed data.

Nonparametric identification, established in (1) and (2), leads to identification of $\psi_0 = (\varphi_0^{\mathrm{T}}, \phi_0^{\mathrm{T}})^{\mathrm{T}}$ under Assumption 2:

$$\varphi_0 = \arg\min_{\varphi\in\Theta_1} \mathrm{E}[S\{\mu_1(X, S=1) - \mu_0(X, S=1) - \tau_\varphi(X)\}^2], \tag{4}$$

$$\phi_0 = \arg\min_{\phi\in\Theta_2} \mathrm{E}[(1-S)\{\mu_1(X, S=0) - \mu_0(X, S=0) - \lambda_\phi(X) - \tau_{\varphi_0}(X)\}^2], \tag{5}$$

and φ_0 and ϕ_0 are the unique values that satisfy (4) and (5). With model misspecification, $\tau_{\varphi_0}(X)$ and $\lambda_{\phi_0}(X)$ can be interpreted as the best approximations of $\tau(X)$ and $\lambda(X)$ in the overlap population (Li, Morgan and Zaslavsky, 2018); see Remark 2. A substantial body of literature has advocated for the use of parametric structural models and projection-based interpretations (e.g., Neugebauer and van der Laan, 2005, Chernozhukov et al., 2018a, Kennedy, Lorch and Small, 2019, Kennedy, Balakrishnan and Wasserman, 2023).

2.4. Direct estimators and the need for improved estimators

Proposition 1 gives two identification strategies and motivates two direct estimators. To construct the direct estimators, let the adjusted outcomes be $Y_i^{\mathrm{adj},1} = \widehat{\mu}_1(X_i, S_i = 1) - \widehat{\mu}_0(X_i; S_i = 1)\}$ and $Y_i^{\mathrm{adj},2} = Y_i\{A_i - \widehat{e}(X_i, S_i)\}/[\widehat{e}(X_i, S_i)\{1 - \widehat{e}(X_i, S_i)\}]$, where $\widehat{\mu}_a(X, S)$ and $\widehat{e}(X, S)$ are estimators of $\mu_a(X, S)$ and $e(X, S)$ for $a = 0, 1$. For $k = 1$ or 2, one can then fit the adjusted outcome $Y_i^{\mathrm{adj},k}$ with

mean $\tau_\varphi(X)+(1-S)\lambda_\phi(X)$ to obtain the direct estimators. However, the two direct estimators require either the correct specification of the outcome mean function or the propensity score. One can use flexible semiparametric or nonparametric models to estimate the two nuisance functions; however, the corresponding direct estimators will suffer from a slower rate of convergence due to the slower-rate of convergence of the nuisance function estimators (Chernozhukov et al., 2018b). This calls for the construction of more principled estimators that provide more attractive statistical properties. It is well-known that estimators constructed based on efficient scores are doubly robust in the sense that they are consistent if either one of the parametric models for the nuisance functions is correctly specified (Robins, Rotnitzky and Zhao, 1994). More recently, many authors have shown that doubly robust estimators possess a "rate-double robustness" property (Rotnitzky, Smucler and Robins, 2021) in the sense that they retain a root-N convergence rate under weaker conditions on consistent but otherwise flexible semi-/non-parametric models of the nuisance functions (Chernozhukov et al., 2018b). In the next section, we derive the semiparametric efficiency score for $\tau_\varphi(X)$ and $\lambda_\phi(X)$ to motivate a new estimator.

3. Semiparametric efficiency theory for $\tau_{\varphi_0}(X)$ and $\lambda_{\phi_0}(X)$

3.1. Semiparametric models with conditional moment restrictions

Our semiparametric model consists of structural models (3), Assumption 1, and other unspecified components of the likelihood function. We show that Assumption 1 imposes restrictions on the structural parameters of interest.

For the trial participants ($S=1$), the transportability assumption of $\tau(X)$ and trial design lead to

$$\mathrm{E}[Y \mid A=1, X, S=1] - \mu_0(X, S=1) = \tau(X). \tag{6}$$

Moreover, for the observational participants ($S=0$), the transportability assumption of $\tau(X)$ and definition of $\lambda(X)$ lead to

$$\mathrm{E}[Y \mid A=1, X, S=0] - \mu_0(X, S=0) = \tau(X) + (1-S)\lambda(X). \tag{7}$$

Combining (6) and (7) results in

$$\mathrm{E}[Y \mid A=1, X, S] = \tau(X) + (1-S)\lambda(X) + \mu_0(X, S). \tag{8}$$

Based on (8), the key insight is to introduce

$$H_{\psi_0} = Y - \{\tau_{\varphi_0}(X) + (1-S)\lambda_{\phi_0}(X)\}A, \tag{9}$$

which enjoys a mean exchangeability property of $\mathrm{E}[H_{\psi_0} \mid A, X, S] = \mu_0(X, S)$ and consequently a conditional moment restriction.

Proposition 2 (Conditional moment restriction). *Under Assumptions 1 and 2, for H_{ψ_0}, we have*

$$\mathrm{E}[H_{\psi_0} \mid A, X, S] = \mathrm{E}[H_{\psi_0} \mid X, S] = \mu_0(X, S). \tag{10}$$

3.2. Semiparametric efficient score

The likelihood function based on a single variable V is $\mathcal{L}(\psi_0, \theta; V) = f(\epsilon_{H,\psi_0} \mid A, X, S) f(A \mid X, S) f(X, S)$, where

$$\epsilon_{H,\psi_0} = H_{\psi_0} - \mathrm{E}[H_{\psi_0} \mid X, S] = Y - \mu(X, S) - \{\tau_{\varphi_0}(X) + (1 - S)\lambda_{\phi_0}(X)\}\{A - e(X, S)\}, \tag{11}$$

and θ is a infinite-dimensional nuisance parameter. The general geometric approach of Bickel et al. (1993) to obtaining the efficient score requires deriving the nuisance tangent space Λ of θ and the projection of the score function of ψ_0 onto $\Lambda^{\perp}$, the orthogonal complement space of Λ. This task is nontrivial because Assumption 1 imposes restrictions on ψ_0 by (10) or equivalently $\mathrm{E}[\epsilon_{H,\psi_0} \mid A, X, S] = 0$. To resolve this challenge, following Robins (1994), we will translate the restrictions directly into the observed data likelihood function, leading to an unconstrained likelihood function of ψ_0, see (21), and finally the efficient score $S_{\psi_0}(V)$. A detailed roadmap and relevant propositions are provided in §8.1 to illustrate the derivation.

Theorem 1 (Semiparametric efficient score of ψ_0). *Suppose Assumptions 1 and 2 hold. The efficient score of ψ_0 is*

$$S_{\psi_0}(V) = \begin{pmatrix} \frac{\partial \tau_{\varphi_0}(X)}{\partial \varphi} \\ (1 - S)\frac{\partial \lambda_{\phi_0}(X)}{\partial \phi} \end{pmatrix} \Big(A - \mathrm{E}\left[AW \mid X, S\right] \mathrm{E}[W \mid X, S]^{-1}\Big) W \epsilon_{H,\psi_0}, \tag{12}$$

where $W = \{\sigma_A^2(X, S)\}^{-1}$, and $V_{\mathrm{eff}} = (\mathrm{E}[S_{\psi_0}(V) S_{\psi_0}^{\mathrm{T}}(V)])^{-1}$ is the semiparametric efficiency bound.

Theorem 1 provides a benchmark for gauging the efficiency of estimators of ψ_0. The efficient score in (12) depends on the unknown distribution through the nuisance functions $\vartheta = (e, \mu, \sigma_a^2)$, indicating $e(X, S)$, $\mu(X, S)$, and $\sigma_a^2(X, S)$, respectively. To facilitate the estimation of ψ_0, we approximate the nuisance functions by flexible semiparametric or nonparametric models and solve the estimating equation of ψ_0 with the approximated nuisance functions based on the observed data. The variance function $\sigma_a^2(X, S)$ can be estimated by fitting a model for $\log\{Y_i - \widehat{\mu}_a(X_i, S_i)\}^2$ against X_i, separately for the treatment group and data source, and transforming the fitted models to the exponential scale. In the simulation study, we adopt the super learner (Van der Laan, Polley and Hubbard, 2007), with the candidate learners including generalized linear models, generalized additive models, and multivariate adaptive regression splines, which can be carried out using off-the-shelf software, e.g., the "SuperLearner" function with specified algorithms in R. To emphasize the dependence on the nuisance functions ϑ, we write $S_{\psi}(V)$ in (12) as $S_{\psi}(V; \vartheta)$. The proposed estimator $\widehat{\psi} = (\widehat{\varphi}^{\mathrm{T}}, \widehat{\phi}^{\mathrm{T}})^{\mathrm{T}}$ solves $\mathrm{P}_N S_{\psi}(V; \widehat{\vartheta}) = 0$ with $e(X, S)$, $\mu(X, S)$, and $\sigma_a^2(X, S)$ replaced by their estimators $\widehat{e}(X, S)$, $\widehat{\mu}(X, S)$, and $\widehat{\sigma}_a^2(X, S)$, respectively.

The decomposition (11) has been utilized in different contexts, including partially linear models (Robinson, 1988, Chernozhukov et al., 2018b), structural nested mean models (Robins, 2004, Yang, 2022), causal random forest (Athey, Tibshirani and Wager, 2019), and R-learner of the HTE (Nie and Wager, 2021). This particular formulation leads to a desirable statistical property of $S_{\psi_0}(V)$ known as Neyman orthogonality, which in turn results in the rate-double robustness property of $\widehat{\psi}$ concerning $\widehat{e}(X, S)$ and $\widehat{\mu}(X, S)$. In the next section, we will delve further into studying this property.

4. Asymptotic property

4.1. Robustness to slower rates for nuisance functions

To protect the estimator from model misspecification, suppose $\widehat{\vartheta} = (\widehat{e}, \widehat{\mu}, \widehat{\sigma}_a^2)$ include general semi-/non-parametric estimators $\widehat{e}(X,S)$, $\widehat{\mu}(X,S)$, and $\widehat{\sigma}_a^2(X,S)$. Let ϑ_0 be the probability limit of $\widehat{\vartheta}$. For a vector v, we use $\|v\|_2 = (v^\top v)^{1/2}$ to denote its Euclidean norm. For a function $g(V)$, denote $\mathrm{P}g(V) = \int g(v)\mathrm{dP}(v)$ and the L_2-norm $\|g(V)\| = \left\{\int g(v)^2 \mathrm{dP}(v)\right\}^{1/2}$. For simplicity of the exposition, we assume the sample size ratio $n/m \to p \in (0,1)$, as $n \to \infty$, so that the asymptotic regime is the same for $n \to \infty$, $m \to \infty$ or $N = n + m \to \infty$. We discuss the case with $n/m \to 0$ in Remark 1. Suppose that $\|\widehat{e}(X,S) - e(X,S)\| = o_{\mathrm{P}}(N^{-\alpha_e})$ and $\|\widehat{\mu}(X,S) - \mu(X,S)\| = o_{\mathrm{P}}(N^{-\alpha_\mu})$. The following theorem summarizes the regularity conditions and asymptotic properties of $\widehat{\psi}$. The proof is relegated to §8.2.

Theorem 2 (Rate-double robustness). *Suppose the assumptions in Theorem 1 hold. Assume further the following regularity conditions hold:*

Condition 1. *ψ_0 is the unique solution to $\mathrm{P}S_\psi(V;\vartheta_0) = 0$, and for any sequence ψ_n, $||\mathrm{P}S_{\psi_n}(V;\vartheta_0)||_2 \to 0$ implies $||\psi_n - \psi_0||_2 \to 0$.*

Condition 2. *(i) $\mathrm{P}_N \partial S_\psi(V;\vartheta)/\partial\psi$ exists and converges uniformly for ψ and ϑ in the neighborhoods of their true values, and (ii) $\Psi = \mathrm{E}[\partial S_{\psi_0}(V;\vartheta_0)/\partial\psi]$ is non-singular.*

Condition 3. *$S_{\psi_0}(V;\widehat{\vartheta})$ and $S_{\psi_0}(V;\vartheta_0)$ belong to a Donsker class of functions (van der Vaart and Wellner, 1996).*

Condition 4. *$|\partial\tau_{\varphi_0}(X)/\partial\varphi^{\mathrm{T}}|$, $|\partial\lambda_{\phi_0}(X)/\partial\phi^{\mathrm{T}}|$, and $\{\widehat{\sigma}_a^2(X,S)\}^{-1}$ are uniformly bounded.*

Condition 5. *The convergence rates of $\widehat{e}(X,S)$ and $\widehat{\mu}(X,S)$ satisfy $\alpha_e \geq 1/4$ and $\alpha_e + \alpha_\mu \geq 1/2$.*

Then, we have $||\widehat{\psi} - \psi_0|| = o_{\mathrm{P}}(1)$, and

$$N^{1/2}(\widehat{\psi} - \psi_0) \to \mathcal{N}\left\{0, \Sigma_{\Psi_0} = (\Psi^{-1})^{\mathrm{T}}\mathrm{E}[S_{\psi_0}(V;\vartheta_0)^{\otimes 2}]\Psi^{-1}\right\}, \tag{13}$$

in distribution, as $N \to \infty$. Moreover, if $\sum_{a=0}^{1} ||\widehat{\sigma}_a^2(X,S) - \sigma_a^2(X,S)|| = o_{\mathrm{P}}(1)$, the asymptotic variance of $\widehat{\psi}$ achieves the semiparametric efficiency bound V_{eff} in Theorem 1.

We discuss the implications of the regularity conditions. Condition 1 is the identifiability condition. Condition 2 is standard in the Z-estimation literature (van der Vaart, 2000). The Donsker class condition in Condition 3 requires that the nuisance functions should not be too complex without imposing independence between the estimated nuisance functions and the data. We refer the interested readers to §4.2 of Kennedy (2016) for a thorough discussion of Donsker classes of functions. Relaxing this condition is possible by using the sample splitting and cross fitting technique for estimation (Chernozhukov et al., 2018b). See §S7.2 in the supplementary material for technical details and empirical evidence. Condition 5 requires $\widehat{e}(X,S)$ and $\widehat{\mu}(X,S)$ to converge to $e(X,S)$ and $\mu(X,S)$ at the rates that make the remaining term in the empirical process negligible; namely,

$$\begin{aligned} ||\mathrm{P}S_{\psi_0}(V;\widehat{\vartheta})||^2 &\preceq ||\widehat{e}(X,S) - e(X,S)|| \times \{||\widehat{\mu}(X,S) - \mu(X,S)|| + ||\widehat{e}(X,S) - e(X,S)||\} \\ &= o_{\mathrm{P}}(N^{-1/2}), \end{aligned}$$

where "$A \preceq B$" denote that A is bounded by a constant times B. Thus, the convergence rate of $\widehat{e}(X,S)$ should be $o_{\mathrm{P}}(N^{-1/4})$, and the convergence rate of $\widehat{\mu}(X,S)$ combined with that of $\widehat{e}(X,S)$ should be $o_{\mathrm{P}}(N^{-1/2})$. In general, there exist different combinations of convergence rates of $\widehat{e}(X,S)$ and $\widehat{\mu}(X,S)$ that result in a negligible error bound accommodating different smoothness conditions of the underlying true nuisance functions, leading to the "rate-double robustness" of $\widehat{\psi}$. This result differs from the mixed bias property of influence functions in Rotnitzky, Smucler and Robins (2021). However, Condition 5 appears similarly in the R-learner of Nie and Wager (2021) due to the similar residual formulation as mentioned in §3.

Importantly, the consistency and asymptotic normality of $\widehat{\psi}$ do not require $\widehat{\sigma}_a^2(X,S)$ to be consistent for $\sigma_a^2(X,S)$ but the efficiency of $\widehat{\psi}$ does. For variance estimation of $\widehat{\psi}$, we approximate the variance formula in (13) by replacing the analytical components with their estimated counterparts, and the expectations with the empirical averages.

Remark 1. The asymptotic covariance matrix of the HTE estimator $\widehat{\varphi}$, denoted by Σ_{φ_0}, can be obtained from the upper $p_1 \times p_1$ block of Σ_{ψ_0}. This matrix allows us to investigate the impact of the distribution of S on $\widehat{\psi}$ and the potential efficiency gain due to data integration. Toward this end, define $r_A^2 = \left(A - \mathrm{E}\left[AW \mid X, S\right] \mathrm{E}[W \mid X, S]^{-1}\right)^2 W$,

$$\Gamma_{1,\mathrm{rct}} = \mathrm{E}\left[S\left\{\frac{\partial \tau_{\varphi_0}(X)}{\partial \varphi}\right\}^{\otimes 2} r_A^2\right], \qquad \Gamma_{12} = \mathrm{E}\left[(1-S)\frac{\partial \tau_{\varphi_0}(X)}{\partial \varphi}\frac{\partial \lambda_{\phi_0}(X)}{\partial \phi^{\mathrm{T}}} r_A^2\right],$$

$$\Gamma_{1,\mathrm{rwd}} = \mathrm{E}\left[(1-S)\left\{\frac{\partial \tau_{\varphi_0}(X)}{\partial \varphi}\right\}^{\otimes 2} r_A^2\right], \qquad \Gamma_2 = \mathrm{E}\left[(1-S)\left\{\frac{\partial \lambda_{\phi_0}(X)}{\partial \phi}\right\}^{\otimes 2} r_A^2\right].$$

Using these notations and some algebra (§S3 of the supplementary material), we have $\Sigma_{\varphi_0} = \left(\Gamma_{1,\mathrm{rct}} + \Gamma_{1,\mathrm{rwd}} - \Gamma_{12}\Gamma_2^{-1}\Gamma_{12}^{\mathrm{T}}\right)^{-1}$. We then obtain a general result that holds even when $n/m \to 0$:

$$N^{1/2}(\Gamma_{1,\mathrm{rct}} + \Gamma_{1,\mathrm{rwd}} - \Gamma_{12}\Gamma_2^{-1}\Gamma_{12}^{\mathrm{T}})^{1/2}(\widehat{\varphi} - \varphi_0) \to \mathcal{N}(0, I_{p_1 \times p_1}),$$

where $I_{p_1 \times p_1}$ is a $p_1 \times p_1$ identity matrix.

The result sheds light on the advantages of using observational studies for possible efficiency gains in treatment effect estimation. The components $\Gamma_{1,\mathrm{rct}}$ and $\Gamma_{1,\mathrm{rwd}} - \Gamma_{12}\Gamma_2^{-1}\Gamma_{12}^{\mathrm{T}}$ in the precision matrix of $\widehat{\varphi}$ depend on the trial sample ($S = 1$) and the observational sample ($S = 0$), respectively. If $\Gamma_{1,\mathrm{rwd}} - \Gamma_{12}\Gamma_2^{-1}\Gamma_{12}^{\mathrm{T}}$ is nonzero, the precision of $\widehat{\varphi}$ is improved by using the observational sample. The next subsection establishes the conditions for achieving efficiency gains through the use of observational studies.

4.2. Efficiency gain of the treatment effect estimation by using the observational studies

We now discuss the advantages of data integration. The trial data grant a consistent estimator of φ_0. Under Assumptions 1 and 2, following the same strategy in §3 for the trial sample, the efficient score of φ_0 is $S_{\mathrm{rct},\varphi_0}(V;\vartheta) = S\{\partial \tau_{\varphi_0}(X)/\partial \varphi\}(A - \mathrm{E}\left[AW \mid X, S\right] \mathrm{E}[W \mid X, S]^{-1})W\epsilon_{H,\psi_0}$. Then, the trial estimator $\widehat{\varphi}_{\mathrm{rct}}$ can be obtained by solving $P_N S_{\mathrm{rct},\varphi}(V;\widehat{\vartheta}) = 0$.

Theorem 3 shows that combining trial and observational studies has the advantage of gaining efficiency in the estimation of φ_0.

Theorem 3 (Efficiency gain by combining trial and observational studies). *Suppose the assumptions in Theorem 2 hold. The asymptotic variance of $\widehat{\varphi}$ is equal to or less than the asymptotic variance of $\widehat{\varphi}_{\text{rct}}$, where the equality holds if*

$$\frac{\partial \tau_{\varphi_0}(X_i)}{\partial \varphi} = M \frac{\partial \lambda_{\phi_0}(X_i)}{\partial \phi} \tag{14}$$

for some constant matrix M. Moreover, the gain in the asymptotic precision, i.e., the inverse of the asymptotic variance, is

$$\{\text{var}_a(\widehat{\varphi})\}^{-1} - \{\text{var}_a(\widehat{\varphi}_{\text{rct}})\}^{-1} = m \times \left(\Omega_{\varphi\varphi} - \Omega_{\varphi\phi}\Omega_{\phi\phi}^{-1}\Omega_{\varphi\phi}^{\text{T}}\right) \geq 0, \tag{15}$$

where var_a *denotes the asymptotic variance, Ω_{ab} is a covariance matrix for $a, b \in \{\varphi, \phi\}$, and recalling m is the sample size of the observational study.*

Exact expressions of Ω_{ab} for $a, b \in \{\varphi, \phi\}$ are provided in the supplementary material. To gain intuition about Theorem 3, it is helpful to discuss two scenarios. If $\lambda_{\phi_0}(X)$ is known, $S_\psi(V;\vartheta)$ uses the additional observational data $(1-S)\{\partial\tau_\varphi(X)/\partial\varphi\}(A - \text{E}[AW \mid X, S]\,\text{E}[W \mid X, S]^{-1})\{\sigma_A^2(X,S)\}^{-1}\epsilon_{H,\psi}$ for estimating φ_0, comparing with $S_{\text{rct},\varphi}(V;\vartheta)$; therefore, $\widehat{\varphi}$ gains precision over $\widehat{\varphi}_{\text{rct}}$. Next, because $\lambda_{\phi_0}(X)$ is unknown, the estimation of ϕ_0 and φ_0 competes for the information in the observational study. When (14) holds, the terms in $\lambda_{\phi_0}(X)$ and that in $\tau_{\varphi_0}(X)$ are collinear, and all observational data are used to estimate ϕ_0. In this case, $\widehat{\psi}$ and $\widehat{\psi}_{\text{rct}}$ have the same asymptotic precision. When (14) does not hold, the terms in $\lambda_{\phi_0}(X)$ and that in $\tau_{\varphi_0}(X)$ are not entirely linearly dependent, and the observational data are used to estimate both ϕ_0 and φ_0. In this case, $\widehat{\psi}$ gains precision over $\widehat{\psi}_{\text{rct}}$, the magnitude of the gain increases with the observational sample size.

Remark 2. The integrative framework shows the efficiency benefits of combining the trial and observational samples over using only the trial sample under Assumptions 1 and 2. Drawn on the semiparametric theory, the projection parameter (φ_0, ϕ_0) results from a projection of the structural models on a constrained model space, enjoys good theoretical properties in terms of consistency and asymptotic efficiency, and is identical to the true model parameters when the posited parametric models are correct. If the underpinning assumptions for the observational sample are violated, potential biases may offset the efficiency benefits. Thus, it is important to scrutinize the required assumptions in practice.

When the putative models for $\tau(X)$ and $\lambda(X)$ are misspecified, $\tau_{\varphi_0}(X)$ and $\lambda_{\phi_0}(X)$ are different from the true estimands, and thus the estimators are biased for $\tau(X)$ and $\lambda(X)$. However, $\tau_{\varphi_0}(X)$ and $\lambda_{\phi_0}(X)$ can be interpreted as the best approximations of $\tau(X)$ and $\lambda(X)$ in the sense of

$$(\varphi_0, \phi_0) = \arg\min_{\varphi,\phi} \text{E}_W[\omega(X,S)[\tau(X) - \tau_\varphi(X) + (1-S)\{\lambda(X) - \lambda_\phi(X)]^2],$$

where $\text{E}_W[g(V) \mid X, S] = \text{E}[g(V)W \mid X, S]/\text{E}[W \mid X, S]$ for any $g(V)$ and $\omega(X,S) = \text{E}_W[A \mid X, S](1 - \text{E}_W[A \mid X, S])$ is the overlap weight (Li, Morgan and Zaslavsky, 2018); see a proof in §S2. Additional simulations under model misspecification of $\tau(x)$ and $\lambda(x)$ are provided in §S7.3 of the supplementary material, which confirms the above statement.

In practice, a goodness-of-fit test can also be developed to assess the adequacy of the structural models using over-identification restrictions; see Yang and Lok (2016) and also §S4.

5. Improve average treatment effect estimation

The HTE characterizes individual variations of the treatment effect, while the average treatment effect τ_0 summarizes the treatment effect for the target patient population at large. Because the trial assigns treatments randomly to the participants, $\tau_{\varphi_0}(X)$ is identifiable and can be estimated. However, due to the inclusion and exclusion criteria for recruiting patients, the patient composition in the trial may be different from the target population; i.e., $f(X \mid S=1)$ is different from $f(X)$ in general. Consequently, $\mathrm{E}[\tau_{\varphi_0}(X) \mid S=1]$ is different from τ_0, and the estimator using the trial data only is biased of τ_0 generally. On the other hand, the observational sample is conceivably more representative of the real patient population because of the real-world data collection mechanisms.

Assumption 3. $f(X \mid S=0) = f(X)$.

We allow the support of $f(X \mid S=1)$ and $f(X)$ to be different, and hence we allow the trial sample and the observational sample to have non-overlapping covariate distributions.

A byproduct of the proposed framework is the identification of τ_0.

Proposition 3 (Identification of τ_0). *Under Assumptions 1 and 3, τ_0 is identified by $\tau_0 = \mathrm{E}[\tau(X)] = \mathrm{E}[\tau(X) \mid S=0]$, where $\tau(X)$ is identified by (1).*

The semiparametric efficient score of τ_0 is presented in the following theorem.

Theorem 4. *Suppose Assumptions 1–3 hold. The semiparametric efficient score of τ_0 is*

$$S_{\tau_0}(V) = \frac{1-S}{\pi_0}\{\tau_{\varphi_0}(X) - \tau_0\} + \mathrm{E}\left[\left.\frac{\partial \tau_{\varphi_0}(X)}{\partial \varphi^{\mathrm{T}}}\right| S=0\right] S_{\varphi_0}(V), \tag{16}$$

where $\pi_0 = \mathrm{P}(S=0)$ and $S_{\varphi_0}(V)$ is the efficient score of φ_0, i.e, the first p_1 components of $S_{\psi_0}(V)$ in (12).

Recall that φ_0 is the parameter in the HTE, ϕ_0 is the parameter in the confounding function, and $\psi_0 = (\varphi_0^{\mathrm{T}}, \phi_0^{\mathrm{T}})^{\mathrm{T}}$ is the combined vector of parameters. From (12), $S_{\varphi_0}(V)$ depends on ϕ_0 in general. Although τ_0 depends only on φ_0 in $\tau_{\varphi_0}(X)$, $S_{\tau_0}(V)$ can depend on ϕ_0 through $S_{\varphi_0}(V)$. From Theorem 4, the observational sample not only provides a representative covariate distribution of the target population but also can contribute to the estimation efficiency of τ_0. Under Assumptions 1 and 3, one can derive the nonparametric efficiency score of τ_0. However, this approach solely utilizes the covariate distribution from the observational sample to adjust for the selection bias present in the trial sample, without incorporating outcome data from the observational sample.

Once we obtain $\widehat{\varphi}$, a simple plug-in estimator of τ_0 is $\widehat{\tau} = m^{-1}\sum_{i=1}^{N}(1-S_i)\tau_{\widehat{\varphi}}(X_i)$. The following theorem shows the rate-double robustness and local efficiency of $\widehat{\tau}$.

Theorem 5. *Suppose the assumptions in Theorem 2 and Assumption 3 hold. Then,*

$$N^{1/2}(\widehat{\tau} - \tau_0) \to \mathcal{N}\left(0, V_{\tau_0}\right), \tag{17}$$

in distribution, as $N \to \infty$, where

$$V_{\tau_0} = \frac{1}{\pi_0}\mathrm{var}\left[\tau_{\varphi_0}(X) \mid S=0\right] + \Psi_0^{\mathrm{T}}\Psi^{-1,\mathrm{T}}\mathrm{E}[S_{\psi_0}(V;\vartheta_0)^{\otimes 2}]\Psi^{-1}\Psi_0,$$

and $\Psi_0 = \mathrm{E}[\partial\tau_{\varphi_0}(X)/\partial\varphi \mid S=0]$. Moreover, the asymptotic variance of $\widehat{\tau}$ achieves the semiparametric efficiency bound.

6. Simulation study

We conduct a simulation study to evaluate the finite sample performance of the proposed estimators of the HTE and τ_0. We follow a similar strategy in Kallus, Puli and Shalit (2018) to generate the data. We first generate the trial data with sample size $n = 1000$. For the trial sample, we sample the covariates X from the superpopulation, where $X_j \sim \mathrm{N}(0,1)$ and $j = 1,\ldots,5$, with the sampling probability $p(X)$, where $\mathrm{logit}\{p(X)\} = X_1/2 - X_2$, to characterize a difference between the patient composition in the trial and the target population, and we generate $A \mid (X, S=1) \sim \mathrm{Ber}(0.5)$ and $Y(a) \mid (X, S=1) = a\tau(X) + (X_1 + X_2X_3/4 - X_4X_5/4) + \exp(X_1/4 - X_2/4)\epsilon(a)$, where $\tau(X) = 1 + X_1 + X_1^2 + X_2 + X_2^2$ and $\epsilon(a) \sim \mathrm{N}(0,1)$, for $a = 0, 1$. We then generate the observational data with sample size $m = 5000$. We sample the covariates X directly from the superpopulation, where $X_j \sim \mathrm{N}(0,1)$ and $j = 1,\ldots,5$, and we generate $A \mid (X, S=0) \sim \mathrm{Ber}\{e(X,0)\}$, where $\mathrm{logit}\{e(X,0)\} = -(X_1X_2/3 + X_3X_4/3 - X_5)$, and $Y(a) \mid (X, S=0) = a\tau(X) + (X_1 + X_2X_3/4 - X_4X_5/4) + (X^{\mathrm{T}}\beta)U + \exp(X_1/4 - X_2/4)\epsilon(a)$, where U is a latent variable and $\epsilon(a) \sim \mathrm{N}(0,1)$, for $a = 0, 1$. We generate U according to a pattern mixture model $U \mid (X, A, S=0) \sim \mathrm{N}(A - 1/2, 1)$, and thus the confounding function is $\lambda(X) = \mu_1(X, S=0) - \mu_0(X, S=0) - \tau(X) = X^{\mathrm{T}}\beta$. We consider two settings: Setting 1 with $\beta = 0 \times (1,\ldots,1)^{\mathrm{T}}$, in which U does not confound the relationship between A and Y, and Setting 2 with $\beta = (1,\ldots,1)^{\mathrm{T}}$, such that U is an unmeasured confounder of A and Y. We consider four estimators: i) $\widehat{\varphi}_{\mathrm{rct}}$ using only the randomized controlled trial data, ii) $\widehat{\varphi}_{\mathrm{meta}}$, meta-analysis of the combined trial and observational studies by regressing the inverse probability of treatment weighting adjusted outcome $Y_i^{\mathrm{adj}} = \{\widehat{e}(X_i, S_i)\}^{-1}A_iY_i - \{1 - \widehat{e}(X_i, S_i)\}^{-1}(1 - A_i)Y_i$ on $(1, X_1, X_1^2, X_2, X_2^2)$, iii) $\widehat{\varphi}$, the proposed integrative estimator with the correctly specified variance model estimated by the linear regression, and iv) $\widehat{\varphi}_{\mathrm{v1}}$, the integrative estimator with the misspecified variance model estimated by 1. The rationale for $\widehat{\varphi}_{\mathrm{meta}}$ is that it uses weighting adjusting for measured confounders and in the absence of unmeasured confounders, $\mathrm{E}[Y_i^{\mathrm{adj}} \mid X_i, S_i] = \tau_{\varphi_0}(X_i)$. For all estimators, we estimate the nuisance functions via the super learner, including candidate learners as generalized linear models, generalized additive models, and multivariate adaptive regression splines. The 95% confidence interval is calculated via parametric-t wild bootstrap as $\left(\tau_{\widehat{\varphi}}(x) - c^*\hat{\mathbb{V}}^{1/2}\{\tau_{\widehat{\varphi}}(x)\}, \tau_{\widehat{\varphi}}(x) + c^*\hat{\mathbb{V}}^{1/2}\{\tau_{\widehat{\varphi}}(x)\}\right)$ at some specific values of x, where c^* is the 95% quantile of the bootstrap t-values $\{|T^{*(b)}| : b = 1, \cdots, B\}$ and $T^{*(b)} = \{\tau_{\widehat{\varphi}^{(b)}}(x) - \tau_{\widehat{\varphi}}(x)\}/\hat{\mathbb{V}}^{1/2}\{\tau_{\widehat{\varphi}^{(b)}}(x)\}$ in each bootstrap iteration. We select the bootstrap size $B = 500$.

Table 1 reports results for point estimation for $\tau(x)$ at various values of x and τ_0, where Integrative denotes the integrative estimator $\tau_{\widehat{\varphi}}(x)$ and Integrative0 denotes $\tau_{\widehat{\varphi}_{\mathrm{v1}}}(x)$. In Setting 1 without unmeasured confounding in the observational study, $\tau_{\widehat{\varphi}_{\mathrm{meta}}}(x)$ shows bias due to the use of the flexible modeling strategy to approximate the propensity score. Among all the estimators, $\tau_{\widehat{\varphi}}(x)$ has a smaller variance than $\tau_{\widehat{\varphi}_{\mathrm{meta}}}(x)$ by capitalizing on semiparametric efficiency theory; $\tau_{\widehat{\varphi}}(x)$ has a smaller variance than $\tau_{\widehat{\varphi}_{\mathrm{rct}}}(x)$ by leveraging the confounding function in the observational study. Although $\tau_{\widehat{\varphi}_{\mathrm{v1}}}(x)$ preserves consistency, its variation is larger than $\tau_{\widehat{\varphi}}(x)$, indicating a loss of efficiency due to the variance model misspecification. In Setting 2 with unmeasured confounding in the observational study, $\tau_{\widehat{\varphi}_{\mathrm{meta}}}(x)$ assuming no unmeasured confounding is biased for $\tau(x)$, due to the unmeasured confounding biases in the observational study, $\tau_{\widehat{\varphi}_{\mathrm{rct}}}(x)$, $\tau_{\widehat{\varphi}}(x)$, and $\tau_{\widehat{\varphi}_{\mathrm{v1}}}(x)$ remain unbiased for $\tau(x)$, and $\tau_{\widehat{\varphi}}(x)$ has improved efficiency over $\tau_{\widehat{\varphi}_{\mathrm{rct}}}(x)$ and $\tau_{\widehat{\varphi}_{\mathrm{v1}}}(x)$. From Table 2, the empirical coverage rates for $\tau_{\widehat{\varphi}}(x)$ and $\tau_{\widehat{\varphi}_{\mathrm{v1}}}(x)$ in both settings with and without unmeasured confounding in the observational study are close to the nominal level. For τ_0, $\hat{\tau}_{\mathrm{rct}}$ is biased when using only the trial data, as the covariate distribution in the trial is different from that in the target population. In contrast, the integrative estimators are consistent by leveraging the representativeness of the covariate distribution in the observational sample.

Table 1. Simulation results for point estimation under two settings with and without unmeasured confounding in the observational study, where the biases are scaled by 10^{-2} and the variances are scaled by 10^{-3}.

	Meta		RCT		Integrative		Integrative0			Meta		RCT		Integrative		Integrative0	
	Bias	Var	Bias	Var	Bias	Var	Bias	Var		Bias	Var	Bias	Var	Bias	Var	Bias	Var
Setting 1 (without unmeasured confounding in the observational study)																	
$\tau(-3,0)$	−49	1024	−4	638	4	203	6	262	$\tau(0,-3)$	110	703	4	828	5	227	3	279
$\tau(-1.5,0)$	−26	100	−1	79	−0	45	1	52	$\tau(0,-1.5)$	54	111	1	98	0	50	−0	58
$\tau(1.5,0)$	22	100	1	79	2	45	1	52	$\tau(0,1.5)$	−59	111	2	98	2	50	2	58
$\tau(3,0)$	48	1024	0	638	8	203	7	262	$\tau(0,3)$	−116	703	6	828	8	227	8	279
$\tau(0,0)$	−2	14	0	22	−1	13	−1	14	τ_0	−2	5	60	17	1	13	1	13
Setting 2 (with unmeasured confounding in the observational study)																	
$\tau(-3,0)$	−324	1060	−4	638	4	203	7	266	$\tau(0,-3)$	−95	691	4	828	5	227	3	282
$\tau(-1.5,0)$	−157	107	−1	79	−0	44	1	52	$\tau(0,-1.5)$	−59	108	1	98	0	50	−0	58
$\tau(1.5,0)$	141	107	1	79	2	44	1	52	$\tau(0,1.5)$	79	108	2	98	2	50	2	58
$\tau(3,0)$	272	1060	0	638	8	203	7	266	$\tau(0,3)$	182	691	6	828	8	227	8	282
$\tau(0,0)$	−1	15	0	22	−1	13	−1	14	τ_0	1	6	60	17	1	13	1	13

7. Real data application

We apply the proposed estimators to evaluate the effect of adjuvant chemotherapy for early-stage resected non-small-cell lung cancer using the CALGB 9633 trial data and a large clinical oncology observational database – the national cancer database. The CALGB 9633 trial used a set of patient eligibility criteria, including disease stage, age, performance status, resection methods, to enroll patients. In the trial sample, 319 patients were randomly assigned to observation versus chemotherapy, resulting 163 on observation, $A = 0$, and 156 on chemotherapy, $A = 1$. The same set of patient eligibility criteria defines the target patient population and is used to select the comparable patients in the national cancer database. The comparable observational sample consists of 15166 patients diagnosed

Table 2. Simulation results for variance estimation and coverage rate for the integrative estimator under two settings with and without unmeasured confounding in the observational study, where the variances are scaled by 10^{-3} and the coverage rates are scaled by 10^{-2}.

	Integrative		Integrative0			Integrative		Integrative0	
	Var	CVG	Var	CVG		Var	CVG	Var	CVG
Setting 1 (without unmeasured confounding in the observational study)									
$\tau(-3,0)$	203	93.3	262	93.1	$\tau(0,-3)$	227	94.4	279	93.9
$\tau(-1.5,0)$	45	94.8	52	94.3	$\tau(0,-1.5)$	50	95.1	58	94.7
$\tau(1.5,0)$	45	93.2	52	94.4	$\tau(0,1.5)$	50	93.5	58	93.4
$\tau(3,0)$	203	94.1	262	93.1	$\tau(0,3)$	227	93.5	279	93.4
$\tau(0,0)$	13	92.6	14	94.2	τ_0	13	93.9	13	93.6
Setting 2 (with unmeasured confounding in the observational study)									
$\tau(-3,0)$	203	93.0	266	92.8	$\tau(0,-3)$	227	94.7	282	94.3
$\tau(-1.5,0)$	44	95.0	52	94.1	$\tau(0,-1.5)$	50	95.2	58	94.6
$\tau(1.5,0)$	44	93.6	52	94.6	$\tau(0,1.5)$	50	93.1	58	93.3
$\tau(3,0)$	203	94.0	266	93.0	$\tau(0,3)$	227	92.9	282	93.0
$\tau(0,0)$	13	93.8	14	93.8	τ_0	13	93.6	13	93.6

Table 3. Sample sizes and covariate means by A in the trial and observational samples.

	N	A	Age (years)	Tumor Size (cm)	Sex (1 = Male) (1/0)	Histology (1 = Squamous) (1/0)	Race (1 = White) (1/0)
Trial	156	$A = 1$	60.6	4.62	64.1%	40.4%	90.4%
	163	$A = 0$	61.1	4.57	63.8%	39.3%	88.3%
Observational study	4263	$A = 1$	63.9	5.19	54.3%	35.6%	88.6%
	10903	$A = 0$	69.4	4.67	54.8%	40.5%	90.0%

with the same disease between years 2004 – 2016 in stage IB disease with 10903 on observation and 4263 received chemotherapy after surgery. As the treatments for the trial patients were randomly assigned and the treatments for the observational patients were chosen by physicians and patients, the numbers of treated and controls are relatively balanced in the trial sample while they are unbalanced in the observational sample. The outcome Y is the indicator of cancer recurrence within three years after the surgery.

We are interested in estimating the heterogeneous treatment effects of chemotherapy varying by tumor size. The original trial analysis did not show any clinical improvement for chemotherapy, possibly because of its small sample size (Strauss et al., 2008). Some exploratory analysis, however, showed that tumor size might modify the treatment effect and that patients with larger tumor sizes may benefit more from the chemotherapy (Strauss et al., 2008, Speicher et al., 2015, Morgensztern et al., 2016). Thus, we formulate the HTE of interest to be $\tau_\varphi(X) = \varphi_1 + \varphi_2 \text{tumor size}^* + \varphi_3(\text{tumor size}^*)^2$, where tumor size* standardizes tumor size by subtracting the mean 4.8 and dividing the standard error 1.7, and $\varphi = (\varphi_1, \varphi_2, \varphi_3)^{\mathrm{T}}$. In the analysis, we include five covariates to adjust for in both samples: age, tumor size, sex, histology, and race, and we use generalized additive models for approximating the nuisance functions. Table 3 reports the covariate means by treatment group in the two samples. Due to treatment randomization, all covariates are balanced between the treated and the control in the trial sample. While due to a lack of treatment randomization, some covariates are highly unbalanced in the observational sample. It can be seen that older patients with smaller tumor sizes and histology are likely to choose a conservative treatment, on observation. Moreover, we cannot rule out the possibility of unmeasured confounders in the observational sample. To formulate the confounding function, possible unmeasured confounders include disease status at diagnosis, financial status, and accessibility to health care facilities that affect the decision of receiving adjuvant chemotherapy after surgery and clinical outcomes (Speicher et al., 2015, Yang et al., 2016, Morgensztern et al., 2016, Speicher et al., 2017). The linear confounding function $\lambda_\phi(X)$ includes age, tumor size, gender, race, histology, Charlson co-morbidity score, income level, insurance coverage, and travel range to large health care facilities as predictors.

We compare the trial, Meta, and integrative estimators. Figure 2 displays the estimated treatment effect as a function of tumor size*. Table 3 reports the results for Tian et al. (2014) the estimated parameters. Due to the small sample size, the trial estimator is not statistically significant. By pooling all information from the trial and observational sample, the Meta and integrative estimators gain efficiency and both show that the tumor size is a significant treatment effect modifier. Interestingly, the two combining approaches produce different conclusions. The difference between the Meta and integrative estimators may be attributable to the no unmeasured confounding. The Meta estimators assume that there are no unmeasured confounders, while the integrative estimators take into account the possible unmeasured confounders in the observational sample. The results in the supplementary material show that age, gender and histology are significant in the confounding function, suggesting that the no unmeasured

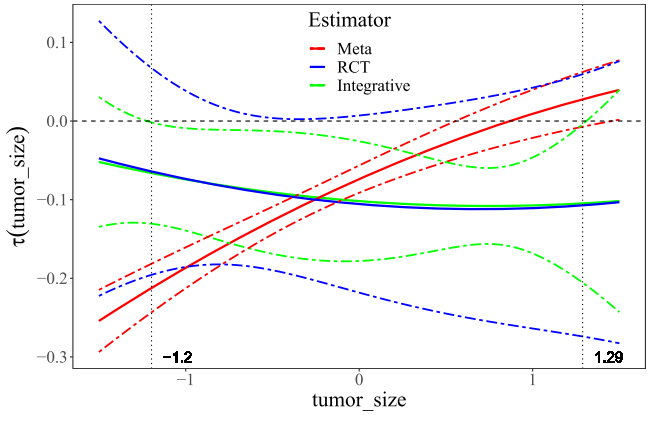

Figure 2. Estimated treatment effect as a function of tumor size*. The solid lines represent the estimators, and the bands represent the solid lines ±1.96 standard errors of the estimators.

confounding assumption is not plausible in the observational sample. For the integrative estimator, we carry out the over-identification restrictions test to assess the goodness-of-fit of $\tau_\psi(X)$ and $\lambda_\phi(X)$. The test is directed at the alternative model specifications $\tau_\psi^{\text{alt}}(X)$ = a quadratic function of age* and tumor size*, and $\lambda_\phi^{\text{alt}}(X) = \lambda_\phi(X)$ augmented with (tumor size*)^2. The test statistics is 5.5 with p-value 0.14 based on a χ_3^2 null reference distribution. Therefore, there is no strong evidence to reject the model specifications of $\tau_\psi(X)$ and $\lambda_\phi(X)$ in this application. From the integrative approach, chemotherapy has significant benefits for patients with tumor size in $[-0.71, 1.2] \times 1.7 + 4.8 = [3.6, 6.8]$cm.

8. Proofs of the main results

As Assumptions 1 and 2 induce the conditional moment restriction, it becomes crucial to take into account such constraint when obtaining the semiparametric efficient score and assessing the asymptotic properties of the integrative estimator. In this section, we provide a detailed derivation of $S_{\psi_0}(V)$ in Theorem 1 and the rate-double robustness in Theorem 2. Those explorations establish the main theoretical contribution of the paper.

8.1. Proof of Theorem 1

We present a roadmap and Propositions 4–6, whose proofs have been included in the supplementary material, to facilitate the construction of the semiparametric efficient score $S_{\psi_0}(V)$ in Theorem 1.

8.1.1. A roadmap

We consider a regular asymptotically linear (RAL) estimator $\widehat{\psi}$ of ψ_0:

$$N^{1/2}(\widehat{\psi} - \psi_0) = N^{-1/2}\mathrm{P}_N \mathrm{IF}(V) + o_{\mathrm{P}}(1), \tag{18}$$

where $\mathrm{IF}(V)$ is the influence function of $\widehat{\psi}$, which has zero mean and finite and nonsingular variance. By (18), the asymptotic variance of $N^{1/2}(\widehat{\psi} - \psi_0)$ is equal to the variance of $\mathrm{IF}(V)$. Consider the Hilbert

space $\mathcal{H}$ of all p-dimensional, mean-zero finite variance squared integrable functions of V, $h(V)$, equipped with the covariance inner product $< h_1, h_2 >= \mathrm{E}[h_1(V)^{\mathrm{T}} h_2(V)]$ and the $\mathcal{L}_2$-norm $||h||^2 = \mathrm{E}[h(V)^{\mathrm{T}} h(V)] < \infty$. To construct the efficient estimator for ψ_0, we follow the geometric approach of Bickel et al. (1993) to derive the semiparametric efficient score for ψ_0 following the road map below.

The density function of a single variable $V = (A, X, Y, S)$ is $f(V) = f(Y \mid A, X, S) f(A \mid X, S) f(X, S)$. The parameter of interest ψ_0 satisfies restriction (10) with $H = H_{\psi_0}$, and the nuisance parameter is the nonparametric density functions $f(Y \mid A, X, S)$, $f(A \mid X, S)$, and $f(X, S)$. In order to incorporate restriction (10) into the likelihood function directly, we consider an equivalent re-parameterization, and re-express the semiparametric likelihood function; see (21). Based on the likelihood function, we characterize the nuisance tangent space Λ in the Hilbert space $\mathcal{H}$; see Proposition 4. We then express Λ as a direct sum of orthogonal subspaces; see Proposition 5. This effort will be valuable in characterizing the orthogonal complement space of the nuisance tangent space $\Lambda^{\perp}$, which consists of all influence functions; see Proposition 6. The semiparametric efficient score of ψ_0 is thus derived as the projection of the score of ψ_0 onto $\Lambda^{\perp}$; see Theorem 1.

8.1.2. Re-parameterization of likelihood function

We consider an equivalent re-parameterization, in order to incorporate restriction (10) into the likelihood function directly. Toward that end, we decompose H as follows:

$$H = \underbrace{H - \mathrm{E}[H \mid A, X, S]}_{\epsilon_H = \epsilon_H(H, A, X, S)} + \underbrace{\mathrm{E}[H \mid A, X, S] - \mathrm{E}[H]}_{Q = Q(A, X, S)} + \mathrm{E}[H], \tag{19}$$

where $\mathrm{E}[\epsilon_H \mid A, X, S] = 0$, $\mathrm{E}[Q] = 0$, and ϵ_H and Q are squared integrable. Note that "squared integrable" is a technical condition to ensure that the nuisance score vectors lie in the Hilbert space $\mathcal{H}$. Then, the semiparametric model defined by restriction (10) is equivalent to the following re-parameterization

$$H = \epsilon_H + q(X, S) - \mathrm{E}[q(X, S)] + \mathrm{E}[H],\ \mathrm{E}[\epsilon_H \mid A, X, S] = 0. \tag{20}$$

On the one hand, if restriction (10) holds, it implies Q depends only on (X, S), but not on A. Because $\mathrm{E}[Q] = 0$, we can then express $Q = q(X, S) - \mathrm{E}[q(X, S)]$ with $q(X, S)$ a squared integrable function of (X, S), so the re-parameterization (20) exists. On the other hand, if H can be expressed in (20), H satisfies the restriction (10).

We can write the likelihood function based on a single variable V as

$$\begin{aligned} \mathcal{L}(\psi, \theta; V) &= f(Y \mid A, X, S) f(A \mid X, S) f(X, S) \\ &= f(\epsilon_H \mid A, X, S) f(A \mid X, S) f(X, S) \frac{\partial \epsilon_H}{\partial Y} \\ &= f(\epsilon_H \mid A, X, S) f(A \mid X, S) f(X, S), \end{aligned} \tag{21}$$

where the last equality follows from

$$\begin{aligned} \epsilon_H = Y &- \{\tau_{\varphi_0}(X) + (1 - S)\lambda_{\phi_0}(X)\} A \\ &- \mathrm{E}[H] - \left\{ q(X, S) - \int q(X, S) f(X, S) \mathrm{d}\nu(X, S) \right\}, \end{aligned}$$

with $q(X, S)$ a nonparametric function of (X, S). Because $\mathrm{E}[\epsilon_H \mid A, X, S] = 0$, we require $\int \epsilon_H f(\epsilon_H \mid A, X, S) \mathrm{d}\nu(\epsilon_H) = 0$, where $\nu(\cdot)$ is a generic measure. After re-parameterization, the nuisance parameter

becomes the infinite dimensional set θ consisting of $f(\epsilon_H \mid A, X, S)$, $f(A \mid X, S)$, $f(X, S)$, $\mathrm{E}[H]$, and $q(X, S)$.

We assume all the regularity conditions to ensure the existence of the efficient score function of ψ_0 are satisfied, which are mainly continuity conditions for the parameter and the semiparametric model; e.g., we need $\psi = \psi(\theta)$ to be pathwise differentiable with respect to θ (Bickel et al., 1993, Tsiatis, 2006). These conditions are not restrictive for a typical application problem.

To distinguish nuisance parameters, we re-write the likelihood function as

$$\mathcal{L}(\psi_0, \theta; V) = f_1(\epsilon_H \mid A, X, S) f_5(A \mid X, S) f_3(X, S), \tag{22}$$

where

$$\epsilon_H = Y - \{\tau_{\varphi_0}(X) + (1 - S)\lambda_{\phi_0}(X)\}A - c_4 - \left\{q_2(X, S) - \int q_2(X, S) f_3(X, S) \mathrm{d}\nu(X, S)\right\},$$

and $\theta = (\theta_1, \ldots, \theta_5)$ consists of the nuisance parameters $\theta_1 = f_1(\epsilon_H \mid A, X, S)$, $\theta_2 = q_2(X, S)$, $\theta_3 = f_3(X, S)$, $\theta_4 = c_4$, and $\theta_5 = f_5(A \mid X, S)$. Then, $\epsilon_H = \epsilon_H(\psi_0, \theta_2, \theta_3, \theta_4)$ depends on the parameter of interest ψ_0 and the nuisance parameters $(\theta_2, \theta_3, \theta_4)$. This order for indexing the nuisance parameters makes the characterization of the nuisance tangent space easier.

Propositions 4 and 5 present the characterizations of the nuisance tangent space and its orthogonal complement, respectively. The proofs are presented in §S1 of the supplementary material.

Proposition 4. *The nuisance tangent space corresponding to* $\theta = (\theta_1, \ldots, \theta_5)$ *is*

$$\Lambda = \Lambda^{(1)} + \Lambda^{(2)} + \Lambda^{(3)} + \Lambda^{(4)} + \Lambda^{(5)},$$

where $\Lambda^{(j)}$ *is the nuisance tangent space with respect to* θ_j*, for* $j = 1, \ldots, 5$*. Define* $\Lambda^* = \{\Gamma^* = \Gamma^*(X, S) \in \mathcal{R}^p : \mathrm{E}[\Gamma^*] = 0\}$ *and* $S_\epsilon = S_\epsilon(\epsilon_H, A, X, S) = \partial \log f_1(\epsilon_H \mid A, X, S)/\partial \epsilon_H \in \mathcal{R}^1$ *evaluated at the truth. Then,*

$$\begin{aligned}
\Lambda^{(1)} &= \{\Gamma^{(1)} = \Gamma^{(1)}(\epsilon_H, A, X, S) \in \mathcal{R}^p : \mathrm{E}[\Gamma^{(1)} \mid A, X, S] = 0, \text{ and } \mathrm{E}[\Gamma^{(1)}\epsilon_H \mid A, X, S] = 0\},\\
\Lambda^{(2)} &= \{\Gamma^{(2)} = \Gamma^{(2)}(\epsilon_H, A, X, S) = \Gamma^{(2)}(\Gamma^*) = \Gamma^* S_\epsilon \in \mathcal{R}^p : \Gamma^* \in \Lambda^*\},\\
\Lambda^{(3)} &= \{\Gamma^{(3)} = \Gamma^{(3)}(\epsilon_H, A, X, S) = \Gamma^{(3)}(\Gamma^*) = \Gamma^* + \mathrm{E}[Q\Gamma^*] S_\epsilon \in \mathcal{R}^p : \Gamma^* \in \Lambda^*\},\\
\Lambda^{(4)} &= \{cS_\epsilon : S_\epsilon = S_\epsilon(\epsilon_H, A, X, S), c \in \mathcal{R}^p\},\\
\Lambda^{(5)} &= \left\{\Gamma^{(5)} = \Gamma^{(5)}(A, X, S) \in \mathcal{R}^p : \mathrm{E}[\Gamma^{(5)}(A, X, S) \mid X, S] = 0\right\}.
\end{aligned}$$

Here and throughout in a slight abuse of notation, we use $\Gamma^{(2)}(\cdot)$ *and* $\Gamma^{(3)}(\cdot)$ *as functions of* (ϵ_H, A, X, S) *and also as operators on* Γ^**, but their meaning should be clear in the context.*

Remark 3. It is important to note that $\Gamma^{(5)}(A, X, S)$ with $\mathrm{E}[\Gamma^{(5)}(A, X, S) \mid X, S] = 0$ is orthogonal to all other subspaces in Λ.

For simplicity, we define the following notation.

Definition 1. Let

$$W = W(A, X, S) = (\mathrm{var}[\epsilon_H \mid A, X, S])^{-1}, \tag{23}$$

$$T = T(X, S) = \mathrm{E}[W \mid X, S], \tag{24}$$

$$\epsilon_0 = \epsilon_0(\epsilon_H, A, X, S) = \mathrm{E}[W \mid X, S]^{-1} W \epsilon_H + Q, \tag{25}$$

$$T^* = \mathrm{E}[T^{-1}] = \mathrm{E}[\mathrm{E}[W \mid X, S]^{-1}]. \tag{26}$$

We now express Λ as a direct sum of orthogonal subspaces. This effort will be valuable in characterizing the orthogonal complement space of the nuisance tangent space $\Lambda^{\perp}$.

Proposition 5. *The space Λ can be written as a direct sum of orthogonal subspaces:*

$$\Lambda = \widetilde{\Lambda}^{(1)} \oplus \widetilde{\Lambda}^{(2)} \oplus \widetilde{\Lambda}^{(3)} \oplus \widetilde{\Lambda}^{(4)} \oplus \widetilde{\Lambda}^{(5)}, \tag{27}$$

where $\oplus$ denotes a direct sum, and using the notation in Proposition 4 and Definition 1, $\widetilde{\Lambda}^{(1)} = \Lambda^{(1)}$,

$$\widetilde{\Lambda}^{(2)} = \left\{ \widetilde{\Gamma}^{(2)} = \widetilde{\Gamma}^{(2)}(\Gamma^*) = \Gamma^* W \epsilon_H : \Gamma^* \in \Lambda^* \right\}, \tag{28}$$

$$\widetilde{\Lambda}^{(3)} = \left\{ \widetilde{\Gamma}^{(3)} = \widetilde{\Gamma}^{(3)}(\Gamma^*) = \Gamma^* - \mathrm{E}[Q\Gamma^*](T^*T)^{-1} W \epsilon_H : \Gamma^* \in \Lambda^* \right\}, \tag{29}$$

$$\widetilde{\Lambda}^{(4)} = \left\{ \widetilde{\Gamma}^{(4)} = c\epsilon_0 : c \in \mathcal{R}^p \right\}, \tag{30}$$

$$\widetilde{\Lambda}^{(5)} = \left\{ \widetilde{\Gamma}^{(5)} = \Gamma^{(5)}(A, X, S) : \mathrm{E}[\Gamma^{(5)}(A, X, S) \mid X, S] = 0 \right\}. \tag{31}$$

Proposition 6. *Suppose Assumptions 1 and 2 hold. The space of the influence function space of ψ_0 is*

$$\Lambda^{\perp} = \left\{ G(A, X, S; \psi_0, c) = c(A, X, S)\epsilon_{H,\psi_0} : \mathrm{E}[c(A, X, S) \mid X, S] = 0 \right\}. \tag{32}$$

8.1.3. Proof of Theorem 1

Based on Proposition 6, we show that the projection of any $B \in \mathcal{H}$, $\prod \left[B \mid \Lambda^{\perp} \right]$, is of the form $c(A, X, S)\epsilon_{H,\psi_0}$, where $\mathrm{E}[c(A, X, S) \mid X, S] = 0$. Let the score vector of ψ_0 be $s_{\psi_0}(V)$. Then, the semiparametric efficient score is the projection of $s_{\psi_0}(V)$ onto $\Lambda^{\perp}$, given by

$$\begin{aligned} S_{\psi_0}(V) = \prod \left[s_{\psi_0}(V) \mid \Lambda^{\perp} \right] = \big(& \mathrm{E}[s_{\psi_0}(V)\epsilon_{H,\psi_0} \mid A, X, S] \\ & - \mathrm{E}[\mathrm{E}[s_{\psi_0}(V)\epsilon_{H,\psi_0} \mid A, X, S] W \mid X, S] \mathrm{E}[W \mid X, S]^{-1} \big) W \epsilon_{H,\psi_0} := c^*(A, X, S)\epsilon_{H,\psi_0}. \end{aligned}$$

To evaluate $c^*(A, X, S)$ further, we note that $\mathrm{E}[\epsilon_{H,\psi_0} \mid A, X, S] = 0$. We differentiate this equality with respect to ψ_0. By the generalized information equality (Newey, 1990), we have $\mathrm{E}[-\partial\epsilon_{H,\psi_0}/\partial\psi \mid A, X, S] + \mathrm{E}[s_{\psi_0}(V)\epsilon_{H,\psi_0} \mid A, X, S] = 0$. Therefore, ignoring the negative sign, we have $c^*(A, X, S)$ as given by

$$\begin{aligned} c^*(A, X, S) &= \left(\mathrm{E}\left[\frac{\partial \epsilon_{H,\psi_0}}{\partial \psi^{\mathrm{T}}} \mid A, X, S \right] - \mathrm{E}\left[\frac{\partial \epsilon_{H,\psi_0}}{\partial \psi^{\mathrm{T}}} W \mid X, S \right] \mathrm{E}[W \mid X, S]^{-1} \right) W \\ &= \begin{pmatrix} \dfrac{\partial \tau_{\varphi_0}(X)}{\partial \varphi} \\ (1 - S)\dfrac{\partial \lambda_{\phi_0}(X)}{\partial \phi} \end{pmatrix} \left(A - \mathrm{E}\left[AW \mid X, S \right] \mathrm{E}[W \mid X, S]^{-1} \right) W. \end{aligned}$$

8.2. Proof of Theorem 2

8.2.1. Preliminaries

We introduce more notations and useful results to prepare for the proof of Theorem 2. Let "$\rightsquigarrow$" denote weak convergence, and let "$A \preceq B$" denote that A is bounded by a constant times B. Denote $\dot{S}_\psi(V;\vartheta) = \partial S_\psi(V;\vartheta)/\partial\psi$. Denote a set of nuisance functions as $\mathcal{G}_{\vartheta_0} = \{\vartheta : ||\vartheta - \vartheta_0|| < \delta\}$ for some $\delta > 0$ and denote $l^\infty(\mathcal{G}_{\vartheta_0})$ as the collection of all bounded functions $f : \mathcal{G}_{\vartheta_0} \to \mathcal{R}^p$.

The following lemmas show the asymptotic properties of functions belong to Donsker classes.

Lemma 1. *Suppose Conditions 2 and 3 hold. Then, we have* $\sup_{\psi\in\Theta,\vartheta\in\mathcal{G}_{\vartheta_0}} ||\mathrm{P}_N S_\psi(V;\vartheta) - \mathrm{P}S_\psi(V;\vartheta)||_2 \to 0$ *in probability as* $N \to \infty$, *and* $\sup_{\psi\in\Theta,\vartheta\in\mathcal{G}_{\vartheta_0}} ||\mathrm{P}_N \dot{S}_\psi(V;\vartheta) - \mathrm{P}\dot{S}_\psi(V;\vartheta)||_2 \to 0$ *in probability as* $N \to \infty$.

Lemma 2. *Suppose Conditions 2 and 3 hold. Then, we have*

$$N^{1/2}(\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta) \rightsquigarrow Z \in l^\infty(\mathcal{G}_{\vartheta_0}),$$

where the limiting process $Z = \{Z(\vartheta) : \vartheta \in \mathcal{G}_{\vartheta_0}\}$ *is a mean-zero multivariate Gaussian process, and the sample paths of* Z *belong to* $\{z \in l^\infty(\mathcal{G}_{\vartheta_0}) : z$ *is uniformly continuous with respect to* $||\cdot||\}$.

8.2.2. Proof of Theorem 3

First, we show the consistency of $\widehat{\psi}$. Toward this end, we show $||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0)||_2 \to 0$. We bound $||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0)||_2$ by

$$\begin{aligned}
||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0)||_2 &\le ||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0) - \mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 + ||\mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 \\
&= ||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0) - \mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 + ||\mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta}) - \mathrm{P}_N S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 \\
&\le ||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0) - \mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 + \sup_{\psi\in\Theta,\vartheta\in\mathcal{G}_{\vartheta_0}} ||\mathrm{P}_N S_\psi(V;\vartheta) - \mathrm{P}S_\psi(V;\vartheta)||_2. \quad (33)
\end{aligned}$$

Both terms in (33) are $o_\mathrm{P}(1)$ as shown below. By the Taylor expansion, we have

$$\begin{aligned}
||S_{\widehat{\psi}}(V;\vartheta_0) - S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 &= \left\| \left.\frac{\partial S_\psi(V;\vartheta)}{\partial\psi^\mathrm{T}}\right|_{\psi=\widehat{\psi},\vartheta=\widetilde{\vartheta}} (\widehat{\vartheta} - \vartheta_0) \right\|_2 \\
&\le \left\| \left.\frac{\partial S_\psi(V;\vartheta)}{\partial\psi^\mathrm{T}}\right|_{\psi=\widehat{\psi},\vartheta=\widetilde{\vartheta}} \right\|_2 \times ||\widehat{\vartheta} - \vartheta_0||_2,
\end{aligned}$$

where $\widetilde{\vartheta}$ lies in the segment between $\widehat{\vartheta}$ and ϑ_0. By the Cauchy–Schwartz inequality, we have

$$\begin{aligned}
||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0) - \mathrm{P}S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 &\le \mathrm{P}||S_{\widehat{\psi}}(V;\vartheta_0) - S_{\widehat{\psi}}(V;\widehat{\vartheta})||_2 \\
&\le \mathrm{P}\left\{ \left\| \left.\frac{\partial S_{\widehat{\psi}}(V;\vartheta)}{\partial\vartheta^\mathrm{T}}\right|_{\vartheta=\widetilde{\vartheta}} \right\|_2 \times ||\widehat{\vartheta} - \vartheta_0||_2 \right\}
\end{aligned}$$

$$\preceq \left\{ \mathrm{E}\left\| \frac{\partial S_{\widehat{\psi}}(V;\vartheta)}{\partial \vartheta^{\mathrm{T}}}\bigg|_{\vartheta=\widetilde{\vartheta}} \right\|_2^2 \right\}^{1/2} \times \left\{ \mathrm{E}||\widehat{\vartheta}-\vartheta_0||_2^2 \right\}^{1/2}$$

$$\preceq ||\widehat{\vartheta}-\vartheta_0||_2 = o_{\mathrm{P}}(1). \tag{34}$$

By Lemma 1, we have

$$\sup_{\psi\in\Theta,\vartheta\in\mathcal{G}_{\vartheta_0}} ||\mathrm{P}_N S_\psi(V;\vartheta) - \mathrm{P}S_\psi(V;\vartheta)||_2 \to 0 \tag{35}$$

in probability as $N\to\infty$. Plugging (34) and (35) into (33) leads to $||\mathrm{P}S_{\widehat{\psi}}(V;\vartheta_0)||_2 = o_{\mathrm{P}}(1)$. Now, by Condition 1, $||\widehat{\psi}-\psi_0||_2 = o_{\mathrm{P}}(1)$.

Second, we show the asymptotic distribution of $\widehat{\psi}$. By the Taylor expansion of $N^{1/2}\mathrm{P}_N S_{\widehat{\psi}}(V;\widehat{\vartheta}) = 0$, we have

$$0 = N^{1/2}\mathrm{P}_N S_{\psi_0}(V;\widehat{\vartheta}) + \{\mathrm{P}_N \dot{S}_{\widetilde{\psi}}(V;\widehat{\vartheta})\}N^{1/2}(\widehat{\psi}-\psi_0),$$

where $\widetilde{\psi}$ lies in the segment between $\widehat{\psi}$ and ψ_0. By Lemma 1, we have

$$\sup_{\psi\in\Theta,\vartheta\in\mathcal{G}_{\vartheta_0}} ||\mathrm{P}_N \dot{S}_\psi(V;\vartheta) - \mathrm{P}\dot{S}_\psi(V;\vartheta)||_2 \to 0$$

in probability as $N\to\infty$. Because $\widetilde{\psi}\to\psi_0$ and $\widehat{\vartheta}\to\vartheta_0$, we have

$$\mathrm{P}_N \dot{S}_{\widetilde{\psi}}(V;\widehat{\vartheta}) \to \Psi = \mathrm{P}\dot{S}_{\psi_0}(V;\vartheta_0)$$

in probability as $N\to\infty$. Thus, we have

$$N^{1/2}(\widehat{\psi}-\psi_0) = -\Psi^{-1}N^{1/2}\mathrm{P}_N S_{\psi_0}(V;\widehat{\vartheta}) + o_{\mathrm{P}}(1). \tag{36}$$

We express

$$\mathrm{P}_N S_{\psi_0}(V;\widehat{\vartheta}) = (\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\widehat{\vartheta}) + \mathrm{P}S_{\psi_0}(V;\widehat{\vartheta}), \tag{37}$$

and show that

$$\mathrm{P}S_{\psi_0}(V;\widehat{\vartheta}) = o_{\mathrm{P}}(N^{-1/2}), \tag{38}$$

$$(\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\widehat{\vartheta}) = (\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta_0) + o_{\mathrm{P}}(N^{-1/2}). \tag{39}$$

To show (38), we denote $c(X,S) = (\partial\tau_{\varphi_0}(X)/\partial\varphi^{\mathrm{T}}, (1-S)\partial\lambda_{\phi_0}(X)/\partial\phi^{\mathrm{T}})^{\mathrm{T}}$ for simplicity and evaluate $\mathrm{P}S_{\psi_0}(V;\widehat{\vartheta})$ explicitly as

$$\begin{aligned}
\mathrm{P}S_{\psi_0}(V;\widehat{\vartheta}) &= \mathrm{E}\left[c(X,S)\left(A\{\widehat{\sigma}_A^2(X,S)\}^{-1} - \{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W}\mid X,S]^{-1}\widehat{W}\right)\widehat{\epsilon}_{H,\psi_0}\right] \\
&= \mathrm{E}\left[c(X,S)\left(\{\widehat{\sigma}_1^2(X,S)\}^{-1}e(X,S) - \{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W}\mid X,S]^{-1}\widehat{W}\right)\right. \\
&\quad \left. \times [\mu_0(X,S) - \widehat{\mu}_0(X,S) - (1-S)\lambda_{\phi_0}(X)\{e(X,S)-\widehat{e}(X,S)\}]\right] \\
&= \mathrm{E}\left[c(X,S)\left(\{\widehat{\sigma}_1^2(X,S)\}^{-1}\{e(X,S)-\widehat{e}(X,S)\}\right.\right. \\
&\quad \left.\left. -\{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W}\mid X,S]^{-1}(\widehat{W}-\widehat{\mathrm{E}}[\widehat{W}\mid X,S])\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\times [\mu_0(X,S) - \widehat{\mu}_0(X,S) - (1-S)\lambda_{\phi_0}(X)\{e(X,S) - \widehat{e}(X,S)\}]\Big] \\
&= \mathrm{E}\Big[c(X,S)\Big(\{\widehat{\sigma}_1^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\} \\
&\quad -\{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W} \mid X,S]^{-1}\{\widehat{\sigma}_1^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\} \\
&\quad +\{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W} \mid X,S]^{-1}\{\widehat{\sigma}_0^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\}\Big) \\
&\quad \times [\mu_0(X,S) - \widehat{\mu}_0(X,S) - (1-S)\lambda_{\phi_0}(X)\{e(X,S) - \widehat{e}(X,S)\}]\Big].
\end{aligned}
$$

Applying the Cauchy–Schwartz inequality and Condition 5, we have

$$
\begin{aligned}
&||\mathrm{P}S_{\psi_0}(V;\widehat{\vartheta})||_2 \\
&\le \mathrm{E}\Big[\Big\|c(X,S)\Big(\{\widehat{\sigma}_1^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\} \\
&\quad -\{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W} \mid X,S]^{-1}\{\widehat{\sigma}_1^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\} \\
&\quad +\{\widehat{\sigma}_1^2(X,S)\}^{-1}\widehat{e}(X,S)\widehat{\mathrm{E}}[\widehat{W} \mid X,S]^{-1}\{\widehat{\sigma}_0^2(X,S)\}^{-1}\{e(X,S) - \widehat{e}(X,S)\}\Big) \\
&\quad \times [\mu_0(X,S) - \widehat{\mu}_0(X,S) - (1-S)\lambda_{\phi_0}(X)\{e(X,S) - \widehat{e}(X,S)\}]\Big\|_2\Big]. \\
&\preceq (\mathrm{E}[\{e(X,S) - \widehat{e}(X,S)\}^2] \times \mathrm{E}\left[\{\mu_0(X,S) - \widehat{\mu}_0(X,S)\}^2 + \{e(X,S) - \widehat{e}(X,S)\}^2\right])^{1/2} \\
&= \{||\widehat{\mu}_0(X,S) - \mu_0(X,S)|| \times ||\widehat{e}(X,S) - e(X,S)|| + ||\widehat{e}(X,S) - e(X,S)||^2\} = o_{\mathrm{P}}(N^{-1/2}).
\end{aligned}
$$

To show (39), Lemma 2 leads to

$$
N^{1/2}(\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta) \rightsquigarrow Z \in l^\infty(\mathcal{G}_{\vartheta_0}),
$$

as $N \to \infty$. Combining with the fact that $||\widehat{\vartheta} - \vartheta_0|| = o_{\mathrm{P}}(1)$, we have

$$
\begin{pmatrix} N^{1/2}(\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta) \\ \widehat{\vartheta} \end{pmatrix} \rightsquigarrow \begin{pmatrix} Z \\ \vartheta_0 \end{pmatrix}
$$

in $l^\infty(\mathcal{G}_{\vartheta_0}) \times \mathcal{G}_{\vartheta_0}$ as $N \to \infty$. Define a function $s : l^\infty(\mathcal{G}_{\vartheta_0}) \times \mathcal{G}_{\vartheta_0} \mapsto \mathcal{R}^p$ by $s(z,\vartheta) = z(\vartheta) - z(\vartheta_0)$, which is continuous for all (z,ϑ) where $\vartheta \mapsto z(\vartheta)$ is continuous. By Lemma 2, all sample paths of Z are continuous on $\mathcal{G}_{\vartheta_0}$, and thus, $s(z,\vartheta)$ is continuous for (Z,ϑ). By the Continuous-Mapping Theorem,

$$
s(Z,\widehat{\vartheta}) = (\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\widehat{\vartheta}) - (\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta_0) \rightsquigarrow s(Z,\vartheta_0) = 0.
$$

Thus, (39) holds. Plugging (37)–(39) into (36), we have

$$
\begin{aligned}
N^{1/2}(\widehat{\psi} - \psi_0) &= -\Psi^{-1}N^{1/2}\{(\mathrm{P}_N - \mathrm{P})S_{\psi_0}(V;\vartheta_0)\} + o_{\mathrm{P}}(1). \\
&\to \mathcal{N}\{0, (\Psi^{-1})^{\mathrm{T}}\mathrm{E}[S_{\psi_0}(V;\vartheta_0)^{\otimes 2}]\Psi^{-1}\},
\end{aligned} \tag{40}
$$

in distribution as $N \to \infty$. If $\sum_{a=0}^{1} ||\widehat{\sigma}_a(X,S) - \sigma_a(X,S)|| = o_{\mathrm{P}}(1)$, $S_{\psi_0}(V;\vartheta_0)$ becomes the efficient score $S_{\psi_0}(V)$. Thus, the asymptotic variance in (40) achieves the efficiency bound. This completes the proof of Theorem 2.

Funding

This project is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD007934 totaling $1,674,013 over two years funded by FDA/HHS. It is also supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG06688, totaling $1,565,763 over four years and the National Science Foundation under Award Number SES 2242776, totaling $225,000 over three years. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, the National Institutes of Health, or the U.S. Government.

Supplementary Material

Supplementary material for "Data fusion methods for the heterogeneity of treatment effect and confounding function" (DOI: 10.3150/24-BEJ1835SUPP; .pdf). The supplementary material includes proofs, simulations, and application results.

References

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455. https://doi.org/10.3386/t0136

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. https://doi.org/10.1007/978-0-387-74978-5_18

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., Adimora, A.A., Eron, J.J. and Mugavero, M.J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. Roy. Statist. Soc. Ser. A* **181** 1193–1209. https://doi.org/10.1111/rssa.12357

Chakraborty, B. and Moodie, E.E.M. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. *Statistics for Biology and Health*. New York: Springer. https://doi.org/10.1007/978-1-4614-7428-9

Chen, D., Hall, P. and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.* **39** 1720–1747. https://doi.org/10.1214/11-AOS882

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36** 808–843. https://doi.org/10.1214/009053607000000947

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018b). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Demirer, M., Duflo, E. and Fernandez-Val, I. (2018a). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India technical report. *NBER Macroecon. Annu.*. https://doi.org/10.3386/w24678

Collins, F.S. and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* **372** 793–795. https://doi.org/10.1001/jama.2015.3595

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J. and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statist. Sci.* **39** 165–191. https://doi.org/10.1214/23-sts889

Dahabreh, I.J., Robertson, S.E., Tchetgen, E.J., Stuart, E.A. and Hernán, M.A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694. https://doi.org/10.1111/biom.13009

Kallus, N., Puli, A.M. and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In *Adv Neural Inf Process Syst* **31** 10888–10897. https://doi.org/10.1101/442442

Kasza, J., Wolfe, R. and Schuster, T. (2017). Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *Int. J. Epidemiol.* **46** 1303–1311. https://doi.org/10.1093/ije/dyx023

Kennedy, E.H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*. *ICSA Book Ser. Stat.* 141–167. Cham: Springer. https://doi.org/10.1007/978-3-319-41259-7_8

Kennedy, E.H., Balakrishnan, S. and Wasserman, L.A. (2023). Semiparametric counterfactual density estimation. *Biometrika* **110** 875–896. https://doi.org/10.1093/biomet/asad017

Kennedy, E.H., Lorch, S. and Small, D.S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 121–143. https://doi.org/10.1111/rssb.12300

Kennedy, E.H., Balakrishnan, S., Robins, J.M. and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *Ann. Statist.* **52** 793–816. https://doi.org/10.1214/24-aos2369

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101** 423–437. https://doi.org/10.1093/biomet/ast066

Lee, D., Yang, S. and Wang, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *J. Causal Inference* **10** 415–440. https://doi.org/10.1515/jci-2022-0004

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D. and Cai, J. (2023). Improving trial generalizability using observational studies. *Biometrics* **79** 1213–1225. https://doi.org/10.1111/biom.13609

Li, F., Morgan, K.L. and Zaslavsky, A.M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. https://doi.org/10.1080/01621459.2016.1260466

Luedtke, A. and Chung, I. (2024). One-step estimation of differentiable Hilbert-valued parameters. *Ann. Statist.* **52** 1534–1563. https://doi.org/10.1214/24-aos2403

Morgensztern, D., Du, L., Waqar, S.N., Patel, A., Samson, P., Devarakonda, S., Gao, F., Robinson, C.G., Bradley, J. and Baggstrom, M. (2016). Adjuvant chemotherapy for patients with T2N0M0 NSCLC. *J. Thorac. Oncol.* **11** 1729–1735. https://doi.org/10.1016/j.jtho.2016.05.022

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *J. Statist. Plann. Inference* **129** 405–426. https://doi.org/10.1016/j.jspi.2004.06.060

Newey, W.K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5** 99–135. https://doi.org/10.3386/w14376

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. https://doi.org/10.1093/biomet/asaa076

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on 540–547*. IEEE. https://doi.org/10.1109/icdmw.2011.169.

Prentice, R.L., Chlebowski, R.T., Stefanick, M.L., Manson, J.E., Pettinger, M., Hendrix, S.L., Hubbell, F.A., Kooperberg, C., Kuller, L.H. and Lane, D.S. (2008). Estrogen plus progestin therapy and breast cancer in recently postmenopausal women. *Amer. J. Epidemiol.* **167** 1207–1216. https://doi.org/10.1093/aje/kwn420

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. https://doi.org/10.1080/03610929408831393

Robins, J.M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. New York: Springer. https://doi.org/10.1007/978-1-4419-9076-1_11

Robins, J.M., Rotnitzky, A. and Scharfstein, D.O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN, 1997)*. *IMA Vol. Math. Appl.* **116** 1–94. New York: Springer. https://doi.org/10.1007/978-1-4612-1284-3_1

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. https://doi.org/10.1080/03610929408831393

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. https://doi.org/10.2307/1912705

Rotnitzky, A., Smucler, E. and Robins, J.M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108** 231–238. https://doi.org/10.1093/biomet/asaa054

Song, R., Luo, S., Zeng, D., Zhang, H.H., Lu, W. and Li, Z. (2017). Semiparametric single-index model for estimating optimal individualized treatment strategy. *Electron. J. Stat.* **11** 364–384. https://doi.org/10.1214/17-EJS1226

Speicher, P.J., Englum, B.R., Ganapathi, A.M., Mulvihill, M.S., Hartwig, M.G., Onaitis, M.W., D'Amico, T.A. and Berry, M.F. (2015). Adjuvant chemotherapy is associated with improved survival after esophagectomy without induction therapy for node-positive adenocarcinoma. *J. Thorac. Oncol.* **10** 181–188. https://doi.org/10.1097/jto.0000000000000384

Speicher, P.J., Englum, B.R., Ganapathi, A.M., Wang, X., Hartwig, M.G., D'Amico, T.A. and Berry, M.F. (2017). Traveling to a high-volume center is associated with improved survival for patients with esophageal cancer. *Ann. Surg.* **265** 743. https://doi.org/10.1097/sla.0000000000001702

Strauss, G.M., Herndon, J.E., II, M.A.M., Johnstone, D.W., Johnson, E.A., Harpole, D.H., Gillenwater, H.H., Watson, D.M., Sugarbaker, D.J. and Schilsky, R.L. (2008). Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non–small-cell lung cancer: CALGB 9633 with the cancer and leukemia group B, radiation therapy oncology group, and North central cancer treatment group study groups. *J. Clin. Oncol.* **26** 5043–5051. https://doi.org/10.3410/f.1127846.595195

Stuart, E.A., Bradshaw, C.P. and Leaf, P.J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* **16** 475–485. https://doi.org/10.1007/s11121-014-0513-z

Stuart, E.A., Cole, S.R., Bradshaw, C.P. and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Tian, L., Alizadeh, A.A., Gentles, A.J. and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266. https://doi.org/10.3102/1076998612441947

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. New York: Springer. https://doi.org/10.1080/03610929408831393

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist. Sci.* **29** 707–731. https://doi.org/10.1214/14-STS493

van der Laan, M.J., Polley, E.C. and Hubbard, A.E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. https://doi.org/10.2202/1544-6115.1309

van der Vaart, A.W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511802256

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. https://doi.org/10.1007/978-1-4757-2545-2

Verde, P.E. and Ohmann, C. (2015). Combining randomized and non-randomized evidence in clinical research: A review of methods and applications. *Res. Synth. Methods* **6** 45–62. https://doi.org/10.1002/jrsm.1122

Yang, S. (2022). Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations. *Biometrics* **78** 937–949. https://doi.org/10.1111/biom.13471

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *J. Amer. Statist. Assoc.* **115** 1540–1554. https://doi.org/10.1080/01621459.2019.1609973

Yang, S. and Lok, J.J. (2016). A goodness-of-fit test for structural nested mean models. *Biometrika* **103** 734–741. https://doi.org/10.1093/biomet/asw031

Yang, S. and Lok, J.J. (2018). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statist. Sinica* **28** 1703–1723. https://doi.org/10.5705/ss.202016.0133

Yang, S. and Wang, X. (2022). RWD-integrated randomized clinical trial analysis *Biopharm. Rep.* **29** 15–21. https://doi.org/10.2147/clep.s48870.

Yang, C.-F.J., Chan, D.Y., Speicher, P.J., Gulack, B.C., Wang, X., Hartwig, M.G., Onaitis, M.W., Tong, B.C., D'Amico, T.A., Berry, M.F. et al. (2016). Role of adjuvant therapy in a population-based cohort of patients with early-stage small-cell lung cancer. *J. Clin. Oncol.* **34** 1057. https://doi.org/10.21037/jtd.2016.10.58

Yang, S., Gao, C., Zeng, D. and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 575–596. https://doi.org/10.1093/jrsssb/qkad017

Yang, S., Liu, S., Zeng, D. and Wang, X. (2025). Supplement to "Data fusion methods for the heterogeneity of treatment effect and confounding function." https://doi.org/10.3150/24-BEJ1835SUPP

Received May 2023 and revised November 2024