Title

Combining Multiple Observational Data Sources to Estimate Causal Effects.

Permalink

https://escholarship.org/uc/item/04j9j16w

Journal

Journal of the American Statistical Association, 115(531)

ISSN

0162-1459

Authors

Yang, Shu

Ding, Peng

Publication Date

2020

DOI

10.1080/01621459.2019.1609973

Peer reviewed

Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(531): 1540–1554. doi:10.1080/01621459.2019.1609973.

Combining Multiple Observational Data Sources to Estimate Causal Effects

Shu Yang[a], **Peng Ding**[b]

[a]Department of Statistics, North Carolina State University, Raleigh, NC

[b]Department of Statistics, University of California, Berkeley, CA

Abstract

The era of big data has witnessed an increasing availability of multiple data sources for statistical analyses. We consider estimation of causal effects combining big main data with unmeasured confounders and smaller validation data with supplementary information on these confounders. Under the unconfoundedness assumption with completely observed confounders, the smaller validation data allow for constructing consistent estimators for causal effects, but the big main data can only give error-prone estimators in general. However, by leveraging the information in the big main data in a principled way, we can improve the estimation efficiencies yet preserve the consistencies of the initial estimators based solely on the validation data. Our framework applies to asymptotically normal estimators, including the commonly used regression imputation, weighting, and matching estimators, and does not require a correct specification of the model relating the unmeasured confounders to the observed variables. We also propose appropriate bootstrap procedures, which makes our method straightforward to implement using software routines for existing estimators. Supplementary materials for this article are available online.

Keywords

Calibration; Causal inference; Inverse probability weighting; Missing confounder; Two-phase sampling

1. Introduction

Unmeasured confounding is an important and common problem in observational studies. Many methods have been proposed to deal with unmeasured confounding in causal inference, such as sensitivity analyses (e.g., Rosenbaum and Rubin 1983a), instrumental variable approaches (e.g., Angrist, Imbens, and Rubin 1996). However, sensitivity analyses cannot provide point estimation, and valid instrumental variables are often difficult to find in practice. We consider the setting where external validation data provide additional

CONTACT Shu Yang syang24@ncsu.edu Department of Statistics, North Carolina State University, 2311 Stinson Drive Campus Box 8203 Raleigh, NC 27695-8203.

Supplementary materials

The online supplementary material contains technical details and proofs. The R package “Integrative CI” is available at https://github.com/shuyang1987/IntegrativeCI to perform the proposed estimators.

information on unmeasured confounders. To be more precise, the study includes a large main dataset representing the population of interest with unmeasured confounders and a smaller validation dataset with additional information about these confounders.

Our framework covers two common types of studies. First, we have a large main dataset, and then collect more information on unmeasured confounders for a subset of units, for example, using a two-phase sampling design (Neyman 1938; Cochran 2007; Wang et al. 2009). Second, we have a smaller but carefully designed validation dataset with rich covariates, and then link it to a larger main dataset with fewer covariates. The second type of data is now ubiquitous. In the era of big data, extremely large data have become available for research purposes, such as electronic health records, claims databases, disease data registries, census data, and to name a few (e.g., Imbens and Lancaster 1994; Schneeweiss et al. 2005; Chatterjee et al. 2016). Although these datasets might not contain full confounder information that guarantees consistent causal effect estimation, they can be useful to increase efficiencies of statistical analyses.

In causal inference, Stürmer et al. (2005) proposed a propensity score calibration method when the main data contain the outcome and an error-prone propensity score based on partial confounders, and the validation data supplement a gold standard propensity score based on all confounders. Stürmer et al. (2005) then applied a regression calibration technique to correct for the measurement error from the error-prone propensity score. This approach does not require the validation data to contain the outcome variable. However, this approach relies on the *surrogacy property* entailing that the outcome variable is conditionally independent of the error-prone propensity score given the gold standard propensity score and treatment. This surrogacy property is difficult to justify in practice, and its violations can lead to substantial biases (Stürmer et al. 2007; Lunt et al. 2012). Under the Bayesian framework, McCandless, Richardson, and Best (2012) specified a full parametric model of the joint distribution for the main and validation data, and treat the gold standard propensity score as a missing variable in the main data. Antonelli, Zigler, and Dominici (2017) combined the ideas of Bayesian model averaging, confounder selection, and missing data imputation into a single framework in this context. Enders et al. (2018) use simulation to show that multiple imputation is more robust than two-phase logistic regression against misspecification of imputation models. Lin and Chen (2014) developed a two-stage calibration method, which summarizes the confounding information through propensity scores and combines the results from the main and validation data. Their two-stage calibration focuses on the regression context with a correctly specified outcome model. Unfortunately, regression parameters, especially in the logistic regression model used by Lin and Chen (2014), may not be the causal parameters of interest in general (Freedman 2008).

In this article, we propose a general framework to estimate causal effects in the setting where the big main data have unmeasured confounders, but the smaller external validation data provide supplementary information on these confounders. Under the assumption of ignorable treatment assignment, causal effects can be identified and estimated from the validation data, using commonly used estimators, such as regression imputation, (augmented) inverse probability weighting (Horvitz and Thompson 1952; Rosenbaum and Rubin 1983b; Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005; Cao, Tsiatis, and

Davidian 2009), and matching (e.g., Rubin 1973; Rosenbaum 1989; Heckman, Ichimura, and Todd 1997; Hirano, Imbens, and Ridder 2003; Hansen 2004; Rubin 2006; Abadie and Imbens 2006; Stuart 2010; Abadie and Imbens 2016). However, these estimators based solely on the validation data may not be efficient. We leverage the correlation between the initial estimator from the validation data and the error-prone estimator from the main data to improve the efficiency over the initial estimator. This idea is similar to the two-stage calibration in Lin and Chen (2014); however, their method focuses only on regression parameters and requires the validation data to be a simple random sample from the main data. Alternatively, the empirical likelihood is also an attractive approach to combine multiple data sources (Chen and Sitter 1999; Qin 2000; Chen, Sitter, and Wu 2002; Chen, Leung, and Qin 2003). However, the empirical likelihood approach needs sophisticated programming, and its computation can be heavy when data become large. Our method is practically simple, because we only need to compute commonly used estimators that can be easily implemented by existing software routines. Moreover, Lin and Chen (2014) and the empirical likelihood approach can only deal with regular and asymptotically linear (RAL) estimators often formulated by moment conditions, but our framework can also deal with non-RAL estimators, such as matching estimators. We also propose a unified bootstrap procedure based on resampling the linear expansions of the estimators, which is simple to implement and works for both RAL and matching estimators.

Furthermore, we relax the assumption that the validation data are a random sample from the study population of interest. We also link the proposed method to existing methods for missing data, viewing the additional confounders as missing values for units outside of the validation data. In contrast to most existing methods in the missing data literature, the proposed method does not need to specify the missing data model relating the unmeasured confounders with the observed variables.

For simplicity of exposition, we use “IID” for “identically and independently distributed,” $1(\cdot)$ for the indicator function, $\xi^{\otimes 2} = \xi\xi^{\mathrm{T}}$ for a vector or matrix ξ “plim” for the probability limit of a random sequence, and $A_n \cong B_n$ for two random sequences satisfying $A_n = B_n + o_P(n^{-1/2})$ with n being the generic sample size. We relegate all regularity conditions for asymptotic analyses to the online supplementary material.

2. Basic Setup

2.1 Notation: Causal Effect and Two Data Sources

Following Neyman (1923) and Rubin 1974), we use the potential outcomes framework to define causal effects. Suppose that the treatment is a binary variable $A \in \{0,1\}$, with 0 and 1 being the labels for control and active treatments, respectively. For each level of treatment $a \in \{0,1\}$, we assume that there exists a potential outcome $Y(a)$, representing the outcome had the subject, possibly contrary to the fact, been given treatment a. The observed outcome is $Y = Y(A) = AY(1) + (1 - A)Y(0)$. Let a vector of pretreatment covariates be (X, U), where X is observed for all units, but U may not be observed for some units.

Although we can extend our discussion to multiple data sources, for simplicity of exposition, we first consider a study with two data sources. The validation data have observations $\mathcal{O}_2 = \{(A_j, X_j, U_j, Y_j): j \in \mathcal{S}_2\}$ with sample size $n_2 = |\mathcal{S}_2|$. The main data have observations $\mathcal{O}_1 = \{(A_i, X_i, Y_i): i \in \mathcal{S}_1 \backslash \mathcal{S}_2\} \cup \mathcal{O}_2$ with sample size $n_1 = |\mathcal{S}_1|$. In our formulation, we consider the case with $\mathcal{S}_2 \subset \mathcal{S}_1$, and let $\rho = \lim_{n_2 \rightarrow \infty} n_2/n_1 \in [0,1]$. If one has two separate main and validation datasets, the main dataset in our context combines these two datasets. Although the main dataset is larger, that is, $n_1 > n_2$, it does not contain full information on important covariates U. Under a superpopulation model, we assume that $\{A_i, X_i, U_i, Y_i(0), Y_i(1) : i \in \mathcal{S}_1\}$ are IID for all $i \in \mathcal{S}_1$, and therefore the observations in $\mathcal{O}_1$ are also IID. The following assumption links the main and validation data.

Assumption 1.—The index set $\mathcal{S}_2$ for the validation data of size n_2 is a simple random sample from $\mathcal{S}_1$.

Under Assumption 1, $\{A_j, X_j, U_j, Y_j(0), Y_j(1) : j \in \mathcal{S}_2\}$ and the observations in $\mathcal{O}_2$ of the validation data are also IID, respectively. We shall relax Assumption 1 to allow $\mathcal{S}_2$ to be a general probability sample from $\mathcal{S}_1$ in Section 7. But Assumption 1 makes the presentation simpler.

Example 1.—Two-phase sampling design is an example that results in the observed data structure. In a study, some variables (e.g., A, X, and Y) may be relatively cheaper, while some variables (e.g., U) are more expensive to obtain. A two-phase sampling design (Neyman 1938; Cochran 2007; Wang et al. 2009) can reduce the cost of the study: in the first phase, the easy-to-obtain variables are measured for all units, and in the second phase, additional expensive variables are measured for a selected validation sample.

Example 2.—Another example is highly relevant in the era of big data, where one links small data with full information on (A, X, U, Y) to external big data with only (A, X, Y). Chatterjee et al. (2016) recently consider this scenario for parametric regression analyses.

Without loss of generality, we first consider the average causal effect (ACE)

$$\tau = E\{Y(1) - Y(0)\}, \tag{1}$$

and will discuss extensions to other causal estimands in Section 4.1. Because of the IID assumption, we drop the indices i and j in the expectations in (1) and later equations.

In what follows, we define the conditional means of the outcome as

$$\mu_a(X, U) = E(Y \mid A = a, X, U),$$

$$\mu_a(X) = E(Y \mid A = a, X),$$

the conditional variances of the outcome as

$$\sigma_a^2(X, U) = \mathrm{var}(Y \mid A = a, X, U),$$

$$\sigma_a^2(X) = \mathrm{var}(Y \mid A = a, X),$$

the conditional probabilities of the treatment as

$$e(X, U) = P(A = 1 \mid X, U), \quad e(X) = P(A = 1 \mid X).$$

2.2 Identification and Model Assumptions

A fundamental problem in causal inference is that we can observe at most one potential outcome for a unit. Following Rosenbaum and Rubin (1983b), we make the following assumptions to identify causal effects.

Assumption 2 (Ignorability).—$Y(a) \perp\!\!\!\perp A \mid (X, U)$ for $a = 0$ and 1.

Under Assumption 2, the treatment assignment is ignorable in $\mathcal{O}_2$ given (X, U). However, the treatment assignment is only "latent" ignorable in $\mathcal{O}_1 \backslash \mathcal{O}_2$ given X and the latent variable U (Frangakis and Rubin 1999; Jin and Rubin 2008).

Moreover, we require adequate overlap between the treatment and control covariate distributions, quantified by the following assumption on the *propensity score* $e(X, U)$.

Assumption 3 (Overlap).—There exist constants c_1 and c_2 such that with probability 1, $0 < c_1 \leq e(X, U) \leq c_2 < 1$.

Under Assumptions 2 and 3, $P\{A = 1 \mid X, U, Y(1)\} = P\{A = 1 \mid X, U, Y(0)\} = e(X, U)$, and $E\{Y(a) \mid X, U\} = E\{Y(a) \mid A = a, X, U\} = \mu_a(X, U)$. The ACE τ can then be estimated through regression imputation, inverse probability weighting (IPW), augmented inverse probability weighting (AIPW), or matching. See Rosenbaum (2002), Imbens (2004), and Rubin (2006) for surveys of these estimators.

In practice, the outcome distribution and the propensity score are often unknown and therefore need to be modeled and estimated.

Assumption 4 (Outcome model).—The parametric model $\mu_a(X, U; \beta_a)$ is a correct specification for $\mu_a(X, U)$,for $a = 0, 1$; that is, $\mu_a(X, U) = \mu_a(X, U; \beta_a^*)$, where β_a^* is the true model parameter, for $a = 0, 1$.

Assumption 5 (Propensity score model).—The parametric model $e(X, U; \alpha)$ is a correct specification for $e(X, U)$; that is, $e(X, U) = e(X, U; \alpha^*)$, where α^* is the true model parameter.

The consistency of different estimators requires different model assumptions.

3. Methodology and Important Estimators

3.1 Review of Commonly Used Estimators Based on Validation Data

The validation data $\{(A_j, X_j, U_j, Y_j) : j \in \mathcal{S}_2\}$ contain observations of all confounders (X, U). Therefore, under Assumptions 2 and 3, τ is identifiable and can be estimated by some commonly used estimator solely from the validation data, denoted by $\hat{\tau}_2$. Although the main data do not contain the full confounding information, we leverage the information on the common variables (A, X, Y) as in the main data to improve the efficiency of $\hat{\tau}_2$. Before presenting the general theory, we first review important estimators that are widely used in practice.

Let $\mu_a(X, U; \beta_a)$ be a working model for $\mu_a(X, U)$, for $a = 0, 1$, and $e(X, U; \alpha)$ be a working model for $e(X, U)$. We construct consistent estimators $\hat{\beta}_a (a = 0, 1)$ and $\hat{\alpha}$ based on $\mathcal{O}_2$, with probability limits $\beta_a^* (a = 0, 1)$ and α^*, respectively. Under Assumption 4, $\mu_a(X, U; \beta_a^*) = \mu_a(X, U)$, and under Assumption 5, $e(X, U; \alpha^*) = e(X, U)$.

Example 3 (Regression imputation).—The regression imputation estimator is $\hat{\tau}_{\text{reg},2} = n_2^{-1} \sum_{j \in \mathcal{S}_2} \hat{\tau}_{\text{reg},2,j}$, where

$$\hat{\tau}_{\text{reg},2,j} = \mu_1(X_j, U_j; \hat{\beta}_1) - \mu_0(X_j, U_j; \hat{\beta}_0) .$$

$\hat{\tau}_{\text{reg},2}$ is consistent for τ under Assumption 4.

Example 4 (Inverseprobability weighting).—The IPW estimator is $\hat{\tau}_{\text{IPW},2} = n_2^{-1} \sum_{j \in \mathcal{S}_2} \hat{\tau}_{\text{IPW},2,j}$, where

$$\hat{\tau}_{\text{IPW},2,j} = \frac{A_j Y_j}{e(X_j, U_j; \hat{\alpha})} - \frac{(1 - A_j) Y_j}{1 - e(X_j, U_j; \hat{\alpha})} .$$

$\hat{\tau}_{\text{IPW},2}$ is consistent for τ under Assumption 5.

The Horvitz-Thompson-type estimator $\hat{\tau}_{\text{IPW},2}$ has large variability, and is often inferior to the Hajek-type estimator (Hájek 1971). We do not present the Hajek-type estimator because we can improve it by the AIPW estimator below. The AIPW estimator employs both the propensity score and the outcome models.

Example 5 (Augmented inverse probability weighting).—Define the residual outcome as $R_j = Y_j - \mu_1(X_j, U_j; \hat{\beta}_1)$ for treated units and $R_j = Y_j - \mu_0(X_j, U_j; \hat{\beta}_0)$ for control units. The AIPW estimator is $\hat{\tau}_{\text{AIPW},2} = n_2^{-1} \sum_{j \in \mathcal{S}_2} \hat{\tau}_{\text{AIPW},2,j}$, where

$$\hat{\tau}_{\text{AIPW},\,2,\,j} = \frac{A_j R_j}{e\left(X_j, U_j; \hat{\alpha}\right)} + \mu_1\left(X_j, U_j; \hat{\beta}_1\right) - \frac{\left(1 - A_j\right)R_j}{1 - e\left(X_j, U_j; \hat{\alpha}\right)} - \mu_0\left(X_j, U_j; \hat{\beta}_0\right). \quad (2)$$

$\hat{\tau}_{\text{AIPW},\,2}$ is doubly robust in the sense that it is consistent if either Assumption 4 or 5 holds. Moreover, it is locally efficient if both Assumptions 4 and 5 hold (Bang and Robins 2005; Tsiatis 2006; Cao, Tsiatis, and Davidian 2009).

Matching estimators are also widely used in practice. To fix ideas, we consider matching with replacement with the number of matches fixed at M. Matching estimators hinge on imputing the missing potential outcome for each unit. To be precise, for unit j, the potential outcome under A_j is the observed outcome Y_j; the (counterfactual) potential outcome under $1 - A_j$ is not observed but can be imputed by the average of the observed outcomes of the nearest M units with $1 - A_j$. Let these matched units for unit j be indexed by $\mathcal{J}_{d,\,V,\,j}$, where the subscripts d and V denote the dataset $\mathcal{O}_d$ and the matching variable V (e.g., $V = (X, U)$), respectively. Without loss of generality, we use the Euclidean distance to determine neighbors; the discussion applies to other distances (Abadie and Imbens 2006). Let $K_{d,\,V,\,j} = \sum_{l \,\in\, \mathcal{S}_d} 1\left(j \in \mathcal{J}_{d,\,V,\,l}\right)$ be the number of times that unit j is used as a match based on the matching variable V in $\mathcal{O}_d$.

Example 6 (Matching).—Define the imputed potential outcomes as

$$\hat{Y}_j(1) = \begin{cases} M^{-1}\sum_{l \,\in\, \mathcal{J}_{2,\,(X,\,U),\,j}} Y_l & \text{if } A_j = 0, \\ Y_j & \text{if } A_j = 1, \end{cases}$$

$$\hat{Y}_j(0) = \begin{cases} Y_j & \text{if } A_j = 0, \\ M^{-1}\sum_{l \,\in\, \mathcal{J}_{2,\,(X,\,U),\,j}} Y_l & \text{if } A_j = 1. \end{cases}$$

Then the matching estimator of τ is

$$\begin{aligned} \hat{\tau}^{(0)}_{\text{mat},\,2} &= n_2^{-1} \sum_{j \,\in\, \mathcal{S}_2} \{\hat{Y}_j(1) - \hat{Y}_j(0)\} \\ &= n_2^{-1} \sum_{j \,\in\, \mathcal{S}_2} \left(2A_j - 1\right)\left(Y_j - M^{-1} \sum_{l \,\in\, \mathcal{J}_{2,\,(X,\,U),\,j}} Y_l\right). \end{aligned}$$

Abadie and Imbens (2006) obtained the decomposition

$$n_2^{1/2}\left(\hat{\tau}^{(0)}_{\text{mat},\,2} - \tau\right) = B_2 + D_2,$$

where

$$B_2 = n_2^{-1/2}\sum_{j \in \mathcal{S}_2}(2A_j - 1) \times \left[M^{-1}\sum_{l \in \mathcal{J}_{2,(X,U),j}}\left\{\mu_{1-A_j}(X_j, U_j) - \mu_{1-A_j}(X_l, U_l)\right\}\right], \tag{3}$$

$$D_2 = n_2^{-1/2}\sum_{j \in \mathcal{S}_2}\Big[\mu_1(X_j, U_j) - \mu_0(X_j, U_j) - \tau + (2A_j - 1)\{1 + M^{-1}K_{2,(X,U),j}\}\left\{Y_j - \mu_{A_j}(X_j, U_j)\right\}\Big].$$

The difference $\mu_{1-A_j}(X_j, U_j) - \mu_{1-A_j}(X_l, U_l)$ in (3) accounts for the matching discrepancy, and therefore B_2 contributes to the asymptotic bias of the matching estimator. Abadie and Imbens (2006) showed that the matching estimators have nonnegligible biases when the dimension of V is greater than one. Let $\hat{\mu}_a(X, U)$ be an estimator for $\mu_a(X, U)$, obtained either parametrically, for example, by a linear regression estimator, or nonparametrically, for $a = 0, 1$. Abadie and Imbens (2006) proposed a bias-corrected matching estimator

$$\hat{\tau}_{\text{mat},2} = \hat{\tau}^{(0)}_{\text{mat},2} - n_2^{-1/2}\hat{B}_2,$$

where $\hat{B}_2$ is an estimator for B_2 by replacing $\mu_a(X, U)$ with $\hat{\mu}_a(X, U)$.

3.2. A General Strategy

We give a general strategy for efficient estimation of the ACE by utilizing both the main and validation data. In Sections 3.3 and 3.4, we will provide examples to elucidate the proposed strategy with specific estimators.

Although the estimators based on the validation data $\mathcal{O}_2$ are consistent for τ under certain regularity conditions, they are inefficient without using the main data $\mathcal{O}_1$. However, the main data $\mathcal{O}_1$ do not contain important confounders U; if we naively use the estimators in Examples 3–6 with U being empty, then the corresponding estimators can be inconsistent for τ and thus are error-prone in general. Moreover, for robustness consideration, we do not want to impose additional modeling assumptions linking U and (A, X, Y).

Our strategy is straightforward: we apply the same error-prone procedure to both the main and validation data. The key insight is that the difference of the two error-prone estimates is consistent for 0 and can be used to improve efficiency of the initial estimator due to its association with $\hat{\tau}_2$. Let an error-prone estimator of τ from the main data be $\hat{\tau}_{1,\text{ep}}$, which converges to some constant τ_{ep}, not necessarily the same as τ. Applying the same method to the validation data $\{(A_j, X_j, Y_j): j \in \mathcal{S}_2\}$, we can obtain another error-prone estimator $\hat{\tau}_{2,\text{ep}}$. More generally, we can consider τ_{ep} to be an L-dimensional vector of parameters identifiable based on the joint distribution of (A, X, Y), and $\hat{\tau}_{1,\text{ep}}$ and $\hat{\tau}_{2,\text{ep}}$ to be the corresponding estimators from the main and validation data, respectively. For example, $\hat{\tau}_{d,\text{ep}}$ an contain estimators of τ using different methods based on $\mathcal{O}_d$.

We consider a class of estimators satisfying

$$n_2^{1/2}\begin{pmatrix}\hat{\tau}_2-\tau\\ \hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}}\end{pmatrix}\rightarrow\mathcal{N}\left\{0_{L+1},\begin{pmatrix}v_2 & \Gamma^{\mathrm{T}}\\ \Gamma & V\end{pmatrix}\right\}, \tag{4}$$

in distribution, as $n_2 \rightarrow \infty$, which is general enough to include all the estimators reviewed in Examples 3–6. Heuristically, if (4) holds exactly rather than asymptotically, by the multivariate normal theory, we have the following the conditional distribution

$$\begin{aligned}&n_2^{1/2}(\hat{\tau}_2-\tau)\mid n_2^{1/2}(\hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}})\\ &\sim\mathcal{N}\left\{n_2^{1/2}\Gamma^{\mathrm{T}}V^{-1}(\hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}}), v_2-\Gamma^{\mathrm{T}}V^{-1}\Gamma\right\}.\end{aligned}$$

Let $\hat{v}_2$, $\hat{\Gamma}$ and $\hat{V}$ be consistent estimators for v_2, Γ and V. We set $n_2^{1/2}(\hat{\tau}_2-\tau)$ to equal its estimated conditional mean $n_2^{1/2}\hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}(\hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}})$, leading to an estimating equation for τ:

$$n_2^{1/2}(\hat{\tau}_2-\tau)=n_2^{1/2}\hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}(\hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}}).$$

Solving this equation for τ, we obtain the estimator

$$\hat{\tau}=\hat{\tau}_2-\hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}(\hat{\tau}_{2,\mathrm{ep}}-\hat{\tau}_{1,\mathrm{ep}}). \tag{5}$$

Proposition 1.—Under Assumption 1 and certain regularity conditions, if (4) holds, then $\hat{\tau}$ is consistent for τ, and

$$n_2^{1/2}(\hat{\tau}-\tau)\rightarrow\mathcal{N}(0, v_2-\Gamma^{\mathrm{T}}V^{-1}\Gamma), \tag{6}$$

in distribution, as $n_2 \rightarrow \infty$. Given a nonzero Γ, the asymptotic variance, $v_2-\Gamma^{\mathrm{T}}V^{-1}\Gamma$, is smaller than the asymptotic variance of $\hat{\tau}_2, v_2$.

The consistency of $\hat{\tau}$ does not require any component in $\hat{\tau}_{1,\mathrm{ep}}$ and $\hat{\tau}_{2,\mathrm{ep}}$ to correctly estimate τ. That is, these estimators can be error prone. The requirement for the error-prone estimators is minimal, as long as they are consistent for the same (finite) parameters. Under Assumption 1, $\hat{\tau}_{1,\mathrm{ep}}-\hat{\tau}_{2,\mathrm{ep}}$ is consistent for a vector of zeros, as $n_2 \rightarrow \infty$.

We can estimate the asymptotic variance of $\hat{\tau}$ by

$$\hat{v}=(\hat{v}_2-\hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}\hat{\Gamma})/n_2. \tag{7}$$

Remark 1.—We construct the error prone estimators $\hat{\tau}_{1,\mathrm{ep}}$ and $\hat{\tau}_{2,\mathrm{ep}}$ based on $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively. Another intuitive way is to construct $\hat{\tau}_{1,\mathrm{ep}}$ and $\hat{\tau}_{2,\mathrm{ep}}$ based on $\mathcal{O}_1\backslash\mathcal{O}_2$ and $\mathcal{O}_2$, respectively. In general, we can construct the error prone estimators based on different subsets of $\mathcal{O}_1$ and $\mathcal{O}_2$ as long as their difference converges in probability to zero. We show in the supplementary material that our construction maximizes the variance reduction for $\hat{\tau}_2, \Gamma^{\mathrm{T}} V^{-1} \Gamma$, given the procedure of the error prone estimators.

Remark 2.—We can view (5) as the best consistent estimator of $\boldsymbol{\tau}$ among all linear combinations $\{\hat{\tau}_2 + \lambda^{\mathrm{T}}(\hat{\tau}_{2,\mathrm{ep}} - \hat{\tau}_{1,\mathrm{ep}}) : \lambda \in \mathbb{R}^L\}$, in the sense that (5) achieves the minimal asymptotic variance among this class of consistent estimators. Similar ideas appeared in design-optimal regression estimation in survey sampling (Deville and Särndal 1992; Fuller 2009), regression analyses (Chen and Chen 2000; Chen 2002; Wang and Wang 2015), improved prediction in high dimensional datasets (Boonstra, Taylor, and Mukherjee 2012), and meta-analysis (Collaboration 2009). In the supplementary material, we show that the proposed estimator in (5) is the best estimator of $\boldsymbol{\tau}$ among the class of estimators $\{\hat{\tau} = f(\hat{\tau}_2, \hat{\tau}_{1,\mathrm{ep}}, \hat{\tau}_{2,\mathrm{ep}}) : f(x, y, z)$ is a smooth function of (x, y, z), and $\hat{\tau}$ is consistent for $\boldsymbol{\tau}\}$, in the sense that (5) achieves the minimal asymptotic variance among this class.

Remark 3.—The choice of the error-prone estimators will affect the efficiency of $\hat{\tau}$. From (6), for a given $\hat{\tau}_2$, to improve the efficiency of $\hat{\tau}$ with a 1-dimensional error-prone estimator, we would like this estimator to have a small variance V and a large correlation with $\hat{\tau}_2$, Γ. In principle, increasing the dimension of the error-prone estimator would not decrease the asymptotic efficiency gain as shown in the supplementary material. However, it would also increase the complexity of implementation and harm the finite sample properties. To "optimize" the tradeoff, we suggest choosing the error-prone estimator to be the same type as the initial estimator $\hat{\tau}_2$. For example, if $\hat{\tau}_2$ is an AIPW estimator, we can choose $\hat{\tau}_{d,\mathrm{ep}}$ to be an AIPW estimator without using U in a possibly misspecified propensity score model. The simulation in Section 5 confirms that this choice is reasonable.

To close this subsection, we comment on the existing literature and the advantages of our strategy. The proposed estimator $\hat{\tau}$ in (5) utilizes both the main and validation data and improves the efficiency of the estimator based solely on the validation data. In economics, Imbens and Lancaster (1994) proposed to use the generalized method of moments (Hansen 1982) for using the main data which provide moments of the marginal distribution of some economic variables. In survey sampling, calibration is a standard technique to integrate auxiliary information in estimation or handle nonresponse; see, for example, Chen and Chen (2000), Wu and Sitter (2001), Kott (2006), Chang and Kott (2008), and Kim, Kwon, and Paik (2016). An important issue is how to specify optimal calibration equations; see, for example, Deville and Särndal (1992), Robins, Rotnitzky, and Zhao (1994), Wu and Sitter (2001), and Lumley, Shaw, and Dai (2011). Other researchers developed constrained empirical likelihood methods to calibrate auxiliary information from the main data; see, for example, Chen and Sitter (1999), Qin (2000), Chen, Sitter, and Wu (2002), and Chen, Leung, and Qin (2003).

Compared to these methods, the proposed framework is attractive because it is simple to implement which requires only standard software routines for existing methods, and it can deal with estimators that cannot be derived from moment conditions, for example, matching estimators. Moreover, our framework does not require a correct model specification of the relationship between unmeasured covariates U and measured variables (A, X, Y).

3.3 Regular Asymptotically Linear (RAL) Estimators

We first elucidate the proposed method with RAL estimators.

From the validation data, we consider the case when $\hat{\tau}_2 - \tau$ is RAL; that is, it can be asymptotically approximated by a sum of IID random vectors with mean 0:

$$\hat{\tau}_2 - \tau \cong n_2^{-1} \sum_{j \in \mathcal{S}_2} \psi(A_j, X_j, U_j, Y_j), \tag{8}$$

where $\{\psi(A_j, X_j, U_j, Y_j): j \in \mathcal{S}_2\}$ are IID with mean 0. The random vector $\psi(A, X, U, Y)$ is called the influence function of $\hat{\tau}_2$ with $E(\psi) = 0$ and $E(\psi^2) < \infty$ (e.g., Bickel et al. 1993). Regarding regularity conditions, see, for example, Newey (1990).

Let $e(X; \gamma)$ be an error-prone propensity score model for, $e(X)$, and $\mu_a(X; \eta_a)$ be an error-prone outcome regression model for $\mu_a(X)$, for $a = 0, 1$. The corresponding error-prone estimators of the ACE can be obtained from the main data $\mathcal{O}_1$ and the validation data $\mathcal{O}_2$. We consider the case when $\hat{\tau}_{d,\mathrm{ep}}$ is RAL:

$$\hat{\tau}_{d,\mathrm{ep}} - \tau_{\mathrm{ep}} \cong n_d^{-1} \sum_{j \in \mathcal{S}_d} \phi\left(A_j, X_j, Y_j\right), \qquad (d = 1, 2) \tag{9}$$

where $\{\phi(A_j, X_j, Y_j): j \in \mathcal{S}_d\}$ are IID with mean 0.

Theorem 1.—Under certain regularity conditions, (4) holds for the RAL estimators (8) and (9), where $v_2 = \mathrm{var}\{\psi(A, X, U, Y)\}$, $\Gamma = (1 - \rho)\mathrm{cov}\{\psi(A, X, U, Y), \phi(A, X, Y)\}$, and $V = (1 - \rho) \times \mathrm{var}\{\phi(A, X, Y)\}$.

To derive $\hat{\Gamma}$ and $\hat{V}$ for RAL estimators, let $\hat{\phi}_d(A, X, Y)$ and $\hat{\psi}(A, X, U, Y)$ be estimators of $\phi(A, X, Y)$ and $\psi(A, X, U, Y)$ by replacing $E(\cdot)$ with the empirical measure and unknown parameters with their corresponding estimators. Note that the subscript d in $\hat{\phi}_d(A, X, Y)$ indicates that it is obtained based on $\mathcal{O}_d$. Then, we can estimate Γ and V by

$$\begin{aligned}\hat{\Gamma} &= \widehat{\mathrm{cov}}(\hat{\tau}_2, \hat{\tau}_{2,\mathrm{ep}} - \hat{\tau}_{1,\mathrm{ep}}) \\ &= \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2} \sum_{j \in \mathcal{S}_2} \hat{\psi}(A_j, X_j, U_j, Y_j)\hat{\phi}_2(A_j, X_j, Y_j),\end{aligned}$$

$$\hat{V} = \widehat{\mathrm{cov}}(\hat{\tau}_{1,\mathrm{ep}} - \hat{\tau}_{2,\mathrm{ep}}) = \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_1} \sum_{i \in \mathcal{S}_1} \left\{\hat{\phi}_1(A_i, X_i, Y_i)\right\}^{\otimes 2}.$$

Finally, we can obtain the estimator and its variance estimator by (5) and (7), respectively.

The commonly-used RAL estimators include the regression imputation and (augmented) inverse probability weighting estimators. Because the influence functions for $\hat{\tau}_{\mathrm{reg},2}$ and $\hat{\tau}_{\mathrm{IPW},2}$ are standard, we present the details in the supplementary material. Below, we state only the influence function for $\hat{\tau}_{\mathrm{AIPW},2}$.

For the outcome model, let $S_a(A, X, U, Y; \beta_a)$ be the estimating function for β_a, for example,

$$S_a(A, X, U, Y; \beta_a) = \frac{\partial \mu_a(X, U; \beta_a)}{\partial \beta_a}\{Y - \mu_a(X, U; \beta_a)\},$$

for $a = 0, 1$, which is a standard choice for the conditional mean model. For the propensity score model, let $S(A, X, U; \alpha)$ be the estimating function for α, for example,

$$S(A, X, U; \alpha) = \frac{A - e(X, U; \alpha)}{e(X, U; \alpha)\{1 - e(X, U; \alpha)\}} \frac{\partial e(X, U; \alpha)}{\partial \alpha},$$

which is the score function from the likelihood of a binary response model. Moreover, let

$$\begin{aligned}\Sigma_{\alpha\alpha} &= E\left\{S^{\otimes 2}(A, X, U; \alpha)\right\} \\ &= E\left[\frac{1}{e(X, U; \alpha^*)\{1 - e(X, U; \alpha^*)\}}\left\{\frac{\partial e(X, U; \alpha^*)}{\partial \alpha}\right\}^{\otimes 2}\right]\end{aligned}$$

be the Fisher information matrix for α in the propensity score model. In addition, let $\hat{\beta}_a(a = 0, 1)$ and $\hat{\alpha}$ be the estimators solving the corresponding empirical estimating equations based on $\mathcal{O}_2$, with probability limits $\beta_a^*(a = 0, 1)$ and α^*, respectively.

Lemma 1 (Augmented inverse probability weighting).—For simplicity, denote $e_j^* = e(X_j, U_j; \alpha^*), \dot{e}_j^* = \partial e(X_j, U_j; \alpha^*)/\partial \alpha^{\mathrm{T}}, S_j^* = S(A_j, X_j, U_j; \alpha^*), \mu_{aj}^* = \mu_a(X_j, U_j; \beta_a^*), \dot{\mu}_{aj}^* = ,$ $\partial \mu_a(X_j, U_j; \beta_a^*)/\partial \beta_a^{\mathrm{T}}, S_{aj}^* = S_a(A_j, X_j, U_j, Y_j; \beta_a^*)$ and $\dot{S}_{aj}^* = \partial S_a(A_j, X_j, U_j, Y_j; \beta_a^*)/\partial \beta_a^{\mathrm{T}}$ for $a = 0, 1$. Under Assumption 4 or 5, $\hat{\tau}_{\mathrm{AIPW},2}$ has the influence function

$$\begin{aligned}&\psi_{\mathrm{AIPW}}(A_j, X_j, U_j, Y_j) \\ &= \frac{A_j Y_j}{e_j^*} + \left(1 - \frac{A_j}{e_j^*}\right)\mu_{1j}^* \\ &\quad - \frac{(1 - A_j)Y_j}{1 - e_j^*} - \left(1 - \frac{1 - A_j}{1 - e_j^*}\right)\mu_{0j}^* - \tau + H_{\mathrm{AIPW}}\Sigma_{\alpha\alpha}^{-1}S_j^* \\ &\quad + E\left\{\left(1 - \frac{1 - A}{1 - e^*}\right)\dot{\mu}_0^*\right\}\left\{E(\dot{S}_0^*)\right\}^{-1}S_{0j}^*\end{aligned} \qquad (10)$$

$$-E\left\{\left(1-\frac{A}{e^{*}}\right)\dot{\mu}_1^{*}\right\}\left\{E(\dot{S}_1^{*})\right\}^{-1}S_{1j}^{*}, \tag{11}$$

where

$$H_{\mathrm{AIPW}}=E\left[\left\{\frac{A(Y-\mu_1^{*})}{(e^{*})^2}-\frac{(1-A)(Y-\mu_0^{*})}{(1-e^{*})^2}\right\}\dot{e}^{*}\right].$$

Lemma 1 follows from standard asymptotic theory, but as far as we know it has not appeared in the literature. Lunceford and Davidian (2004) suggest a formula without (10) and (11) for ψ_{AIPW}, which, however, works only when both Assumptions 4 and 5 hold. Otherwise, the resulting variance estimator is not consistent if either Assumption 4 or 5 does not hold, as shown by simulation in Funk et al. (2011). The correction terms in (10) and (11) also make the variance estimator doubly robust in the sense that the variance estimator for $\hat{\tau}_{\mathrm{AIPW},2}$ is consistent if either Assumption 4 or 5 holds, not necessarily both.

For error-prone estimators, we can obtain the influence functions similarly. The subtlety is that both the propensity score and outcome models can be misspecified. For simplicity of the presentation, we defer the exact formulas to the online supplementary material.

3.4 Matching Estimators

We then elucidate the proposed method with non-RAL estimators. An important class of non-RAL estimators for the ACE are the matching estimators. The matching estimators are not regular estimators because the functional forms are not smooth due to the fixed numbers of matches (Abadie and Imbens 2008). Continuing with Example 6, Abadie and Imbens 2006) express the bias-corrected matching estimator $\hat{\tau}_{\mathrm{mat},2}$ in a linear form as

$$\hat{\tau}_{\mathrm{mat},2}-\tau\cong n_2^{-1}\sum_{j\in\mathcal{S}_2}\psi_{\mathrm{mat},j}, \tag{12}$$

where

$$\begin{aligned}\psi_{\mathrm{mat},j}=&\ \mu_1(X_j,U_j)-\mu_0(X_j,U_j)-\tau\\&+(2A_j-1)\left\{1+M^{-1}K_{2,(X,U),j}\right\}\left\{Y_j-\mu_{A_j}(X_j,U_j)\right\}.\end{aligned} \tag{13}$$

Similarly, $\hat{\tau}_{\mathrm{mat},d,\mathrm{ep}}$ has a linear form

$$\hat{\tau}_{\mathrm{mat},d,\mathrm{ep}}-\tau_{\mathrm{ep}}\cong n_d^{-1}\sum_{j\in\mathcal{S}_d}\phi_{\mathrm{mat},d,j}, \tag{14}$$

where

$$\begin{aligned}\phi_{\mathrm{mat},d,j}=&\ \mu_1(X_j)-\mu_0(X_j)-\tau_{\mathrm{ep}}\\&+(2A_j-1)(1+M^{-1}K_{d,X,j})\left\{Y_j-\mu_{A_j}(X_j)\right\}.\end{aligned} \tag{15}$$

Theorem 2.—Under certain regularity conditions, (4) holds for the matching estimators (12) and (14), where

$$
\begin{aligned}
v_2 &= \mathrm{var}\left\{\tau(X,U)\right\} \\
&\quad + \mathrm{plim}\left[n_2^{-1}\sum_{j\in\mathcal{S}_2}\left\{1+M^{-1}K_{2,(X,U),j}\right\}^2\sigma^2_{A_j}\left(X_j,U_j\right)\right], \\
\Gamma &= (1-\rho)(\mathrm{cov}\{\mu_1(X,U)-\mu_0(X,U),\mu_1(X)-\mu_0(X)\} \\
&\quad + \mathrm{plim}\left[n_2^{-1}\sum_{j\in\mathcal{S}_2}\left\{1+M^{-1}K_{2,(X,U),j}\right\}\right. \\
&\quad \left.\times\left(1+M^{-1}K_{2,X,j}\right)\sigma^2_{A_j}\left(X_j,U_j\right)\right]),
\end{aligned}
$$

$$
\begin{aligned}
V = (1-\rho)&\left[\mathrm{var}\{\mu_1(X)-\mu_0(X)\}\right. \\
&\left.+\,\mathrm{plim}\left\{n_2^{-1}\sum_{j\in\mathcal{S}_2}\left(1+M^{-1}K_{2,X,j}\right)^2\sigma^2_{A_j}(X_j)\right\}\right].
\end{aligned}
$$

The existence of the probability limits in Theorem 2 are guaranteed by the regularity conditions specified in the supplementary material (c.f. Abadie and Imbens 2006).

To estimate (v_2, Γ, V) in Theorem 2, we need to estimate the conditional mean and variance functions of the outcome given covariates. Following Abadie and Imbens (2006), we can estimate these functions via matching units with the same treatment level. We will discuss an alternative bootstrap strategy in the next subsection.

3.5 Bootstrap Variance Estimation

The asymptotic results in Theorems 1 and 2 allow for variance estimation of $\hat{\tau}$. In addition, we also consider the bootstrap for variance estimation, which is simpler to implement and often has better finite sample performances (Otsu and Rai 2016). This is particularly important for matching estimators because the analytic variance formulas involve nonparametric estimation of the conditional variances $\sigma_a^2(x,u)$ and $\sigma_a^2(x)$.

There are two approaches for obtaining bootstrap observations: (a) the original observations; and (b) the asymptotic linear terms of the proposed estimator. For RAL estimators, bootstrapping the original observations will yield valid variance estimators (Efron and Tibshirani 1986; Shao and Tu 2012). However, for matching estimators, Abadie and Imbens (2008) showed that due to lack of smoothness in their functional form, the bootstrap based on approach (a) does not apply for variance estimation. This is mainly because the bootstrap based on approach (a) cannot preserve the distribution of the numbers of times that the units are used as matches. As a remedy, Otsu and Rai (2016) proposed to construct the bootstrap counterparts by resampling based on approach (b) for the matching estimator.

To unify the notation, let ψ_j indicate $\psi(A_j, X_j, U_j, Y_j)$ for RAL $\hat{\tau}_2$ and $\psi_{\mathrm{mat},\, j}$ for $\hat{\tau}_{\mathrm{mat},\, 2}$ and similar definitions apply to $\phi_{d,\, j}(d = 1, 2)$. Let $\hat{\psi}_j$ and $\hat{\phi}_{d,\, j}$ be their estimated version by replacing the population quantities by the estimated quantities (d = 1, 2). Following Otsu and Rai (2016), for b = 1,...,B, we construct the bootstrap replicates for the proposed estimators as follows:

Step 1. Sample n_1 units from $\mathcal{S}_1$ with replacement as $\mathcal{S}_1^{*(b)}$, treating the units with observed U as the bootstrap validation data $\mathcal{S}_2^{*(b)}$.

Step 2. Compute the bootstrap replicates of $\hat{\tau}_2 - \tau$ and $\hat{\tau}_{d,\,\mathrm{ep}} - \tau_{\mathrm{ep}}$ as

$$\hat{\tau}_2^{(b)} - \hat{\tau}_2 = n_2^{-1} \sum_{j \in S_2^{*}(b)} \hat{\psi}_j,$$

$$\hat{\tau}_{d,\,\mathrm{ep}}^{(b)} - \hat{\tau}_{d,\,\mathrm{ep}} = n_d^{-1} \sum_{j \in \mathcal{S}_d^{*(b)}} \hat{\phi}_{d,\, j}, \qquad (d = 1, 2).$$

Based on the bootstrap replicates, we estimate Γ, V and v_2 by

$$\hat{\Gamma} = (B-1)^{-1} \sum_{b=1}^{B} (\hat{\tau}_2^{(b)} - \hat{\tau}_2)(\hat{\tau}_{2,\,\mathrm{ep}}^{(b)} - \hat{\tau}_{1,\,\mathrm{ep}}^{(b)} - \hat{\tau}_{2,\,\mathrm{ep}} + \hat{\tau}_{1,\,\mathrm{ep}}), \tag{16}$$

$$\hat{V} = (B-1)^{-1} \sum_{b=1}^{B} \left(\hat{\tau}_{2,\,\mathrm{ep}}^{(b)} - \hat{\tau}_{1,\,\mathrm{ep}}^{(b)} - \hat{\tau}_{2,\,\mathrm{ep}} + \hat{\tau}_{1,\,\mathrm{ep}}\right)^{\otimes 2}, \tag{17}$$

$$\hat{v}_2 = (B-1)^{-1} \sum_{b=1}^{B} (\hat{\tau}_2^{(b)} - \hat{\tau}_2)^2 . \tag{18}$$

Finally, we estimate the asymptotic variance of $\hat{\tau}$ by (7), that is, $\hat{v} = (\hat{v}_2 - \hat{\Gamma}^{\mathrm{T}} \hat{V}^{-1} \hat{\Gamma})/n_2$.

Theorem 3.—Under certain regularity conditions, $(\hat{\Gamma}, \hat{V}, \hat{v}_2, \hat{v})$ are consistent for $\{\Gamma, V, \mathrm{var}(\hat{\tau}_2), \mathrm{var}(\hat{\tau})\}$.

Remark 4.—If the ratio of n_2 and n_1 is small, the above bootstrap approach may be unstable, because it is likely that some bootstrap validation data contain only a few or even zero observations. In this case, we use an alternative bootstrap approach, where we sample n_2 units from $\mathcal{S}_2$ with replacement as $\mathcal{S}_2^{*}$, sample $n_1 - n_2$ units from $\mathcal{S}_1 \backslash \mathcal{S}_2$ with replacement, combined with $\mathcal{S}_2^{*}$, as $\mathcal{S}_1^{*}$, and obtain the proposed estimators based on $\mathcal{S}_1^{*}$ and $\mathcal{S}_2^{*}$. This approach guarantees that the bootstrap validation data contain n_2 observations.

Remark 5.—It is worthwhile to comment on a computational issue. When the main data have a substantially large size, the computation for the bootstrap can be demanding if we follow Steps 1 and 2 above. In this case, we can use subsampling (Politis, Romano, and Wolf 1999) or the Bag of Little Bootstraps (Kleiner et al. 2014) to reduce the computational burden. More interestingly, when $n_1 \to \infty$ and $\rho = 0$, that is, the validation data contain a small fraction of the main data, Γ and V reduce to $\mathrm{cov}(\hat{\tau}_2, \hat{\tau}_{2,\mathrm{ep}})$ and $\mathrm{var}(\hat{\tau}_{2,\mathrm{ep}})$, respectively. That is, when the size of the main data is substantially large, we can ignore the uncertainty of $\hat{\tau}_{1,\mathrm{ep}}$ and treat it as a constant, which is a regime recently considered by Chatterjee et al. (2016). In this case, we need only to bootstrap the validation data, which is computationally simpler.

4. Extensions

4.1 Other Causal Estimands

Our strategy extends to a wide class of causal estimands, as long as (4) holds. For example, we can consider the average causal effects over a subset of population (Crump et al. 2006; Li, Morgan, and Zaslavsky 2016), including the average causal effect on the treated.

We can also consider nonlinear causal estimands. For example, for a binary outcome, the log of the causal risk ratio is

$$\log\mathrm{CRR} = \log\frac{P\{Y(1)=1\}}{P\{Y(0)=1\}} = \log\frac{E\{Y(1)\}}{E\{Y(0)\}},$$

and the log of the causal odds ratio is

$$\begin{aligned}\log\mathrm{COR} &= \log\frac{P\{Y(1)=1\}/P\{Y(1)=0\}}{P\{Y(0)=1\}/P\{Y(0)=0\}} \\ &= \log\frac{E\{Y(1)\}/[1-E\{Y(1)\}]}{E\{Y(0)\}/[1-E\{Y(0)\}]}.\end{aligned}$$

We give a brief discussion for the log CRR as an illustration. The key insight is that under Assumptions 2 and 3, we can estimate $E\{Y(a)\}$ with commonly-used estimators from $\mathcal{O}_2$, denoted by $\hat{E}\left\{Y(a)\right\}$, for $a = 0, 1$. We can then obtain an estimator for the log CRR as log $[\hat{E}\left\{Y(1)\right\}/\hat{E}\left\{Y(0)\right\}]$. Similarly, we can obtain error-prone estimators for the log CRR from both $\mathcal{O}_1$ and $\mathcal{O}_2$ using only covariates X. By the Taylor expansion, we can linearize these estimators and establish a similar result as (4), which serves as the basis to construct an improved estimator for the log CRR.

4.2 Design Issue: Optimal Sample Size Allocation

As a design issue, we consider planning a study to obtain the data structure in Example 1 in Section 2 subject to a cost constraint. The goal is to find the optimal design, specifically the sample allocation, that minimizes the variance of the proposed estimator subject to a cost constraint, as in the classical two-phase sampling (Cochran 2007).

Suppose that it costs C_1 to collect (A, X, Y) for each unit, and C_2 to collect U for each unit. Thus, the total cost of the study is

$$C = n_1 C_1 + n_2 C_2 . \tag{19}$$

The variance of the proposed estimator $\hat{\tau}$ is of the form

$$n_2^{-1} v_2 - (n_2^{-1} - n_1^{-1})\gamma, \tag{20}$$

for example, for RAL estimators,

$$\begin{aligned}\gamma = \operatorname{cov}\{\psi(A, X, U, Y), \phi(A, X, Y)\}^{\mathrm{T}}[\operatorname{var}\{\phi(A, X, Y)\}]^{-1} \\ \operatorname{cov}\left\{\psi(A, X, U, Y), \phi(A, X, Y)\right\}\end{aligned}$$

is the variance of the projection of $\psi(A, X, U, Y)$ onto the linear space spanned by $\phi(A, X, Y)$. Minimizing (20) with respect to n_1 and n_2 subject to the constraint (19) yields the optimal n_1^* and n_2^*, which satisfy

$$\rho^* = \frac{n_2^*}{n_1^*} = \left\{(1 - R_{\psi \mid \phi}^2) \times \frac{C_1}{C_2}\right\}^{1/2}, \tag{21}$$

where $R_{\psi \mid \phi}^2 = \gamma / v_2$ is the squared multiple correlation coefficient of $\psi(A, X, U, Y)$ on $\phi(A, X, Y)$, which measures the association between the initial estimator and the error-prone estimator. We derive (21) using the Lagrange multipliers, and relegate the details to the supplementary material. Not surprisingly, (21) shows that the sizes of the validation data and the main data should be inversely proportional to the square-root of the costs. In addition, from (21), a large size n_2 for the validation data is more desirable when the association between the initial estimator and the error-prone estimator is small.

4.3 Multiple Data Sources

We have considered the setting with two data sources, and we can easily extend the theory to the setting with multiple data sources $\mathcal{O}_1, \dots, \mathcal{O}_K$, where $\mathcal{O}_1, \dots, \mathcal{O}_{K-1}$ contain partial covariate information, and the validation data, $\mathcal{O}_K$, contain full information for (A, X, U, Y). For example, for $d = 1, \dots, K - 1$, $\mathcal{O}_d$ contains variables (A, V_d, Y) where $V_d \subsetneqq (X, U)$. Each dataset $\mathcal{O}_d$, indexed by $\mathcal{S}_d$, has size n_d for $d = 1,...,K$. This type of data structure arises from a multi-phase sampling as an extension of Example 1 or multiple sources of “big data” as an extension of Example 2.

Let $\hat{\tau}_K$ be the initial estimator for τ from the validation data $\mathcal{O}_K$, and $\hat{\tau}_{d,\mathrm{ep}}$ be the error-prone estimator for τ from $\mathcal{O}_d (d = 1, \dots, K - 1)$. Let $\hat{\tau}_{d,K,\mathrm{ep}}$ be the estimator obtained by applying the same error-prone estimator for O_d to $\mathcal{O}_K$, so that $\hat{\tau}_{d,\mathrm{ep}} - \hat{\tau}_{d,K,\mathrm{ep}}$ is consistent for 0, for $d = 1, \dots, K - 1$. Assume that

$$n_K^{1/2}\begin{pmatrix}\hat{\tau}_K-\tau\\ \hat{\tau}_{1,\mathrm{ep}}-\hat{\tau}_{1,K,\mathrm{ep}}\\ \vdots\\ \hat{\tau}_{K-1,\mathrm{ep}}-\hat{\tau}_{K-1,K,\mathrm{ep}}\end{pmatrix}$$
$$\to\mathcal{N}\left\{\begin{pmatrix}0\\ 0_L\end{pmatrix},\begin{pmatrix}v_K & \Gamma^{\mathrm{T}}\\ \Gamma & V\end{pmatrix}\right\},$$

in distribution, as $n_K \to \infty$, where $L = \sum_{d=1}^{K-1}\dim(\hat{\tau}_{d,\mathrm{ep}})$. If Γ and V have consistent estimators $\hat{\Gamma}$ and $\hat{V}$, respectively, then, extending the proposed method in Section 3, we can use

$$\hat{\tau} = \hat{\tau}_K - \hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}\left(\hat{\tau}_{1,\mathrm{ep}}^{\mathrm{T}} - \hat{\tau}_{1,K,\mathrm{ep}}^{\mathrm{T}}, \ldots, \hat{\tau}_{K-1,\mathrm{ep}}^{\mathrm{T}} - \hat{\tau}_{K-1,K,\mathrm{ep}}^{\mathrm{T}}\right)^{\mathrm{T}}$$

to estimate $\boldsymbol{\tau}$. The estimator $\hat{\tau}$ is consistent for $\boldsymbol{\tau}$ with the asymptotic variance $v_K - \Gamma^{\mathrm{T}}V^{-1}\Gamma$, which is smaller than the asymptotic variance of $\hat{\tau}_K$, v_K, if Γ is nonzero. Similar to the reasoning in Remark 3, using more data sources will improve the asymptotic estimation efficiency of $\boldsymbol{\tau}$.

5. Simulation

In this section, we conduct a simulation study to evaluate the finite sample performance of the proposed estimators. In our data generating model, the covariates are $X_i \sim \mathrm{Unif}(0,2)$ and $U_i = 0.5 + 0.5X_i - 2\sin(X_i) + 2\mathrm{sign}\{\sin(5X_i)\} + \epsilon_i$, where $\epsilon_i \sim \mathrm{Unif}(-0.5, 0.5)$. The potential outcomes are $Y_i(0) = -X_i - U_i + \epsilon_i(0)$, and $Y_i(1) = -X_i + 4U_i + \epsilon_i(1)$, where $\epsilon_i(0) \sim \mathcal{N}(0,1)$, $\epsilon_i(1) \sim \mathcal{N}(0,1)$, and $\epsilon_i(0)$ and $\epsilon_i(1)$ are independent. Therefore, the true value of the ACE is $\boldsymbol{\tau} = E(5U_i)$. The treatment indicator A_i follows Bernoulli (π_i) with $\mathrm{logit}(\pi_i) = 1 - 0.5X_i - 0.5U_i$. The main data $\mathcal{O}_1$ consist of n_1 units, and the validation data $\mathcal{O}_2$ consist of n_2 units randomly selected from the main data.

The initial estimators are the regression imputation, (A)IPW and matching estimators applied solely to the validation data, denoted by $\hat{\tau}_{\mathrm{reg},2}$, $\hat{\tau}_{\mathrm{IPW},2}$, $\hat{\tau}_{\mathrm{AIPW},2}$, and $\hat{\tau}_{\mathrm{mat},2}$ respectively. To distinguish the estimators constructed based on different error-prone methods, we assign each proposed estimator a name with the form $\hat{\tau}_{\mathrm{method},2\&\mathrm{methods}}$, where "method,2" indicates the initial estimator applied to the validation data $\mathcal{O}_2$ and "methods" indicates the error-prone estimator(s) used to improve the efficiency of the initial estimator. For example, $\hat{\tau}_{\mathrm{reg},2\&\mathrm{IPW}}$ indicates the initial estimator is the regression imputation estimator and the error-prone estimator is the IPW estimator. We compare the proposed estimators with the initial estimators in terms of percentages of reduction of mean squared errors, defined as $\{1 - \mathrm{MSE}(\hat{\tau}_{\mathrm{method},2\&\mathrm{methods}})/\mathrm{MSE}(\hat{\tau}_{\mathrm{method},2})\} \times 100\%$. To demonstrate the robustness of the proposed estimator against misspecification of the imputation model, we consider the multiple imputation (MI, Rubin 1987) estimator, denoted by $\hat{\tau}_{\mathrm{mi}}$, which uses a regression model of U given (A, X, Y) for imputation. We implement MI using the "mice" package in R with $m = 10$.

Based on a point estimate $\hat{\tau}$ and a variance estimate $\hat{v}$ obtained by the asymptotic variance formula or the bootstrap method described in Section 3.5, we construct a Wald-type 95% confidence interval $(\hat{\tau} - z_{0.975}\hat{v}^{1/2}, \hat{\tau} + z_{0.975}\hat{v}^{1/2})$, where $z_{0.975}$ is the 97.5% quantile of the standard normal distribution. We further compare the variance estimators in terms of empirical coverage rates.

Figure 1 shows the simulation results over 2000 Monte Carlo samples for $(n_1, n_2) = (1000,200)$ and $(n_1, n_2) = (1000,500)$. The multiple imputation estimator is biased due to the misspecification of the imputation model. In all scenarios, the proposed estimators are unbiased and improve the initial estimators. Using the error-prone estimator of the same type of the initial estimator achieves a substantial efficiency gain, and the efficiency gain from incorporating additional error-prone estimator is not significantly important. Because of the practical simplicity, we recommend using the same type of error-prone estimator to improve the efficiency of the initial estimator. Confidence intervals constructed from the asymptotic variance formula and the bootstrap method work well, in the sense that the empirical coverage rate of the confidence intervals is close to the nominal coverage rate. In our settings, the matching estimator has the smallest efficiency gain among all types of estimators.

6. Application

We present an analysis to evaluate the effect of chronic obstructive pulmonary disease (COPD) on the development of herpes zoster (HZ). COPD is a chronic inflammatory lung disease that causes obstructed airflow from the lungs, which can cause systematic inflammation and dysregulate a patient's immune function. The hypothesis is that people with COPD are at increased risk of developing HZ. Yang et al. (2011) find a positive association between COPD and development of HZ; however, they do not control for important counfounders between COPD and HZ, for example, cigarette smoking and alcohol consumption.

We analyze the main data from the 2005 Longitudinal Health Insurance Database (LHID, Yang et al. 2011) and the validation data from the 2005 National Health Interview Survey conducted by the National Health Research Institute and the Bureau of Health Promotion in Taiwan (Lin and Chen 2014). The 2005 LHID consist of 42,430 subjects followed from the date of cohort entry on January 1, 2004 until the development of HZ or December 31, 2006, whichever came first. Among those, there are 8,486 subjects with COPD, denoted by $A = 1$, and 33,944 subjects without COPD, denoted by $A = 0$. The outcome Y was the development of HZ during follow up (1, having HZ and 0, not having HZ). The observed prevalence of HZ among COPD and non-COPD subjects are 3.7% and 2.2% in the main data and 2.5% and 0.8% in the validation data.

The confounders X available from the main data were age, sex, diabetes mellitus, hypertension, coronary artery disease, chronic liver disease, autoimmune disease, and cancer. However, important confounders U, including cigarette smoking and alcohol consumption, were not available. The validation data $\mathcal{O}_2$ use the same inclusion criteria as in the main study and consist of 1,148 subjects who were comparable to the subjects in the main data. Among those, 244 subjects were diagnosed of COPD, and 904 subjects were not. In addition to all variables available from the main data, cigarette smoking and alcohol consumption were measured. In our formulation, the main data $\mathcal{O}_1$ combine the LHID data and the validation data. Table 4 in Lin and Chen (2014) shows summary statistics on demographic characteristics and comorbid disorders for COPD and Non-COPD subjects in the main and validation data. Because the common covariates in the main and validation data are comparable, it is reasonable to assume that the validation sample is a simple random sample from the main data. Moreover, the difference in distributions of alcohol consumption between COPD and non-COPD subjects is not statistical significant in the validation data. But, the COPD subjects tended to have higher cumulative smoking rates than the non-COPD subjects in the validation data.

We obtain the initial estimators applied solely to the validation data and the proposed estimators applied to both data. As suggested by the simulation in Section 5, we use the same type of the error-prone estimator as the initial estimator. Following Stürmer et al. (2005) and Lin and Chen (2014), we use the propensity score to accommodate the high-dimensional confounders. Specifically, we fit logistic regression models for the propensity score $e(X, U; \alpha)$ and the error-prone propensity score $e(X; \gamma)$ based on $\{(A_j, X_j, U_j) : j \in \mathcal{S}_2\}$ and $\{(A_i, X_i) : j \in \mathcal{S}_1\}$, respectively. We fit logistic regression models for the outcome mean function $\mu_a(X, U)$ based on a linear predictor $\{1, e(X, U; \hat{\alpha})\}^{\mathrm{T}}\beta_a$, and for $\mu_a(X)$ based on a linear predictor $\{1, e(X; \hat{\gamma})\}^{\mathrm{T}}\eta_a$, for $a = 0, 1$.

We first estimate the ACE τ. Table 1 shows the results for the average COPD effect on the development of HZ. We find no big differences in the point estimates between our proposed estimators and the corresponding initial estimators, but large reductions in the estimated standard errors of the proposed estimators. As a result, all 95% confidence intervals based on the initial estimators include 0, but the 95% confidence intervals based on the proposed estimators do not include 0, except for $\hat{\tau}_{\mathrm{mat2\&mat}}$. As demonstrated by the simulation in Section 5, the variance reduction by utilizing the main data is the smallest for the matching estimator. From the results, on average, COPD increases the percentage of developing HZ by 1.55%.

We also estimate the log of the causal risk ratio of HZ with COPD. The initial IPW estimate from the validation data is $\log \widehat{\mathrm{CRR}}_{\mathrm{IPW},\,2} = 1.10$ (95% confidence interval: 0.02, 2.18). In contrast, the proposed estimate by using the error-prone IPW estimators is log $\widehat{\mathrm{CRR}}_{\mathrm{IPW},\,2\&\mathrm{IPW}} = 0.57$ (95% confidence interval: 0.41, 0.72), which is much more accurate than the initial IPW estimate.

7. Relaxing Assumption 1

In previous sections, we invoked Assumption 1 that $\mathcal{S}_2$ is a random sample from $\mathcal{S}_1$. We now relax this assumption and link our framework to existing methods for missing data. Let I_i be the indicator of selecting unit i into the validation data, that is, $I_i = 1$ if $i \in \mathcal{S}_2$ and $I_i = 0$ if $i \notin \mathcal{S}_2$. Alternatively, I_i can be viewed as the missingness indicator of U_i. Under Assumption 1, $I \perp\!\!\!\perp (A, X, U, Y)$; that is, U is missing completely at random. We now relax it to $I \perp\!\!\!\perp U \mid (A, X, Y)$, that is, U is missing at random. In this case, the selection of $\mathcal{S}_2$ from $\mathcal{S}_1$ can depend on a probability design, which is common in observational studies, for example, an outcome-dependent two-phase sampling (Breslow, McNeney, and Wellner 2003; Wang et al. 2009).

We assume that each unit in the main data is subjected to an independent Bernoulli trial which determines whether the unit is selected into the validation data. For simplicity, we further assume that the inclusion probability $P(I = 1 \mid A, X, U, Y) = P(I = 1 \mid A, X, Y) \equiv \pi(A, X, Y)$ is known as in two-phase sampling. Otherwise, we need to fit a model for the missing data indicator I given (A, X, Y). We summarize the above in the following assumption.

Assumption 6.

$\{(I_i, A_i, X_i, U_i, Y_i):\ i \in \mathcal{S}_1\}$ are IID with $I \perp\!\!\!\perp U \mid (A, X, Y)$. $\mathcal{S}_2$ is selected from $\mathcal{S}_1$ with a known inclusion probability $\pi(A, X, Y) > 0$.

In what follows, we use π for $\pi(A, X, Y)$ and π_j for $\pi(A_j, X_j, Y_j)$ for shorthand. Because of Assumption 6, we drop the indices i and j in the expectations, covariances, and variances, which are taken with respect to both the sampling and superpopulation models.

7.1 RAL estimators

For the illustration of RAL estimators, we focus on the AIPW estimator of the ACE τ, because the regression imputation and inverse probability weighting estimators are its special cases. Let $\hat{\alpha}$ and $\hat{\beta}_a$ solve the weighted estimating equations $\sum_{j \in \mathcal{S}_2} \pi_j^{-1} S(A_j, X_j, U_j; \alpha) = 0$ and $\sum_{j \in \mathcal{S}_2} \pi_j^{-1} S_a\left(A_j, X_j, U_j, Y_j; \beta_a\right) = 0$, and let α^* and β_a^* satisfy $E\{S(A, X, U; \alpha^*)\} = 0$ and $E\{S_a(A, X, U, Y; \beta_a^*)\} = 0$. Under suitable regularity condition, $\hat{\alpha} \to \alpha^*$ and $\hat{\beta}_a \to \beta_a^*$ in probability, for $a = 0, 1$. Let the initial estimator for τ be the Hajek-type estimator (Hájek 1971):

$$\hat{\tau}_2 = \frac{\sum_{j \in \mathcal{S}_2} \pi_j^{-1} \hat{\tau}_{\mathrm{AIPW}, 2, j}}{\sum_{j \in \mathcal{S}_2} \pi_j^{-1}}, \qquad (22)$$

where $\hat{\tau}_{\mathrm{AIPW}, 2, j}$ has the same form as (2). Under regularity conditions, Assumption 4 or 5, and Assumption 6, we show in the supplementary material that

$$\hat{\tau}_2 - \tau \cong n_1^{-1} \sum_{j \in \mathcal{S}_1} \pi_j^{-1} I_j \psi(A_j, X_j, U_j, Y_j), \tag{23}$$

where $\psi(A, X, U, Y)$ is given by (11). Because $\{\pi_j^{-1} I_j \psi(A_j, X_j, U_j, Y_j) : j \in \mathcal{S}_1\}$ are IID with mean 0, $\hat{\tau}_2$ is consistent for τ.

Similarly, let $\hat{\gamma}_d$ and $\hat{\eta}_{d,a}$ solve the weighted estimating equation $\sum_{j \in \mathcal{S}_d} \pi_j^{-1} S(A_j, X_j; \gamma) = 0$ and $\sum_{j \in \mathcal{S}_d} \pi_j^{-1} S_a(A_j, X_j, Y_j; \eta_a) = 0$, and let γ^* and η_a^* satisfy $E\{S(A_j, X_j; \gamma^*)\} = 0$ and $E\{S_a(A_j, X_j, Y_j; \eta_a^*)\} = 0$. Under suitable regularity condition, $\hat{\gamma}_d \to \gamma^*$ and $\hat{\eta}_{d,a} \to \eta_a^*$ in probability, for $a = 0, 1$ and $d = 1, 2$. Let the error-prone estimators be

$$\hat{\tau}_{1,\text{ep}} = n_1^{-1} \sum_{i \in \mathcal{S}_1} \hat{\tau}_{\text{AIPW},1,\text{ep},i}, \quad \hat{\tau}_{2,\text{ep}} = \frac{\sum_{j \in \mathcal{S}_2} \pi_j^{-1} \hat{\tau}_{\text{AIPW},2,\text{ep},j}}{\sum_{j \in \mathcal{S}_2} \pi_j^{-1}}, \tag{24}$$

where $\hat{\tau}_{\text{AIPW},d,\text{ep},j}$ has the same form as (S8) in the supplementary material. Following a similar derivation for (23), we have

$$\begin{aligned} \hat{\tau}_{1,\text{ep}} - \tau_{\text{ep}} &\cong n_1^{-1} \sum_{i \in \mathcal{S}_1} \phi(A_i, X_i, Y_i), \\ \hat{\tau}_{2,\text{ep}} - \tau_{\text{ep}} &\cong n_1^{-1} \sum_{j \in \mathcal{S}_1} \pi_j^{-1} I_j \phi(A_j, X_j, Y_j), \end{aligned} \tag{25}$$

where $\phi(A, X, Y)$ is given by (S9) in the supplementary material. Because both $\{\phi(A_i, X_i, Y_i) : i \in \mathcal{S}_1\}$ and $\left\{\pi_j^{-1} I_j \phi\left(A_j, X_j, Y_j\right) : j \in \mathcal{S}_1\right\}$ are IID with mean 0, $\hat{\tau}_{1,\text{ep}}$ and $\hat{\tau}_{2,\text{ep}}$ are consistent for τ_{ep}.

Theorem 4.—Under certain regularity conditions, (4) holds for the Hajek-type estimators (22) and (24), where $\rho = \text{plim}_{n_2 \to \infty}(n_2/n_1), \quad v_2 = \rho \times \text{var}\{\pi^{-1} I \psi(A, X, U, Y)\}, \Gamma = \rho \times \text{cov}\{\pi^{-1} I \psi(A, X, U, Y), (\pi^{-1} I - 1)\phi(A, X, Y)\}$ and $V = \rho \times \text{var}\{(\pi^{-1} I - 1)\phi(A, X, Y)\}$.

Similar to Section 3.3, we can construct a consistent variance estimator for $\hat{\tau}$ by replacing the variances and covariance in Theorem 4 with their sample analogs.

7.2 Matching Estimators

Recall that $\mathcal{J}_{d,V,l}$ is the index set of matches for unit l based on data $\mathcal{O}_d$ and the matching variable V, which can be (X, U) or X. Define $\delta_{d,V,(j,l)} = 1$ if $j \in \mathcal{J}_{d,V,l}$ and $\delta_{d,V,(j,l)} = 0$ otherwise. Now, we denote $K_{d,V,j} = \pi_j \sum_{l \in \mathcal{S}_d} \pi_l^{-1} 1\{A_l = 1 - A_j\} \delta_{d,V,(j,l)}$ as the weighted number of times that unit j is used as a match. If π_j is a constant for all $j \in \mathcal{S}_d$, then $K_{d,V,j}$ reduces to the number of times that unit j is used as a match defined in Section 3.1, which justifies using the same notation as before.

Let the initial matching estimator for τ be the Hajek-type estimator:

$$
\begin{aligned}
\hat{\tau}^{(0)}_{\mathrm{mat},2} = & \left(\sum_{j\in\mathcal{S}_2}\pi_j^{-1}\right)^{-1} \\
& \times \sum_{j\in\mathcal{S}_2}\pi_j^{-1}(2A_j-1)\left(Y_j - M^{-1}\sum_{l\in\mathcal{J}_{2,(X,U),j}} Y_l\right) \\
= & \left(\sum_{j\in\mathcal{S}_2}\pi_j^{-1}\right)^{-1} \\
& \times \sum_{j\in\mathcal{S}_2}\pi_j^{-1}(2A_j-1)\left\{1 + M^{-1}K_{2,(X,U),j}\right\}Y_j .
\end{aligned}
$$

Let a bias-corrected matching estimator be

$$
\hat{\tau}_{\mathrm{mat},2} = \hat{\tau}^{(0)}_{\mathrm{mat},2} - n_1^{-1/2}\hat{B}_2, \tag{26}
$$

where

$$
\begin{aligned}
\hat{B}_2 = n_1^{-1/2} & \sum_{j\in\mathcal{S}_2}\pi_j^{-1}(2A_j-1) \\
\times & \left[M^{-1}\sum_{l\in\mathcal{J}_{2,(X,U),j}}\left\{\hat{\mu}_{1-A_j}(X_j,U_j) - \hat{\mu}_{1-A_j}(X_l,U_l)\right\}\right],
\end{aligned}
$$

We show in the supplementary material that

$$
\hat{\tau}_{\mathrm{mat},2} - \tau \cong n_1^{-1}\sum_{j\in\mathcal{S}_2}\pi_j^{-1}\psi_{\mathrm{mat},j}, \tag{27}
$$

where $\psi_{\mathrm{mat},j}$ is defined in (13) with the new definition of $K_{2,(X,U),j}$.

Similarly, we obtain error-prone matching estimators and express them as

$$
\begin{aligned}
\hat{\tau}_{\mathrm{mat},1,\mathrm{ep}} - \tau_{\mathrm{ep}} &\cong n_1^{-1}\sum_{j\in\mathcal{S}_1}\phi_{\mathrm{mat},1,j}, \\
\hat{\tau}_{\mathrm{mat},2,\mathrm{ep}} - \tau_{\mathrm{ep}} &\cong n_1^{-1}\sum_{j\in\mathcal{S}_2}\pi_j^{-1}\phi_{\mathrm{mat},2,j},
\end{aligned} \tag{28}
$$

where $\phi_{\mathrm{mat},d,j}$ is defined in (15) with the new definition of $K_{d,X,j}$.

From the above decompositions, $\hat{\tau}_{\mathrm{mat},2}$ is consistent for τ, and $\hat{\tau}_{\mathrm{mat},1,\mathrm{ep}} - \hat{\tau}_{\mathrm{mat},2,\mathrm{ep}}$ is consistent for 0.

Theorem 5.—Under certain regularity conditions, (4) holds for the estimators (26) and $\hat{\tau}_{\mathrm{mat},d,\mathrm{ep}}(d=1,2)$, where $\rho = \mathrm{plim}_{n_2\to\infty}(n_2/n_1)$,

$$\begin{aligned}\upsilon_2 = \rho \times \Bigg(&E\left[\frac{1-\pi}{\pi}\{\tau(X,U)-\tau\}^2\right]\\ &+ \text{plim}\left[n_1^{-1}\sum_{j\in\mathcal{S}_1}\frac{1-\pi_j}{\pi_j}\left\{1+M^{-1}K_{2,(X,U),j}\right\}^2\right.\\ &\left.\sigma^2_{A_j}(X_j,U_j)\right]\Bigg),\end{aligned}$$

$$\begin{aligned}\Gamma = \rho\times E\Bigg[&\frac{1-\pi}{\pi}\{\mu_1(X,U)-\mu_0(X,U)-\tau\}\\ &\left\{\mu_1(X)-\mu_0(X)-\tau_{ep}\right\}\Bigg]\\ &+\rho\times\text{plim}\left[n_1^{-1}\sum_{j\in\mathcal{S}_1}\frac{1-\pi_j}{\pi_j}\left\{1+M^{-1}K_{2,(X,U),j}\right\}\right.\\ &\left.(1+M^{-1}K_{2,X,j})\sigma^2_{A_j}(X_j,U_j)\right]\end{aligned}$$

$$\begin{aligned}V = \rho\times E&\left[\frac{1-\pi}{\pi}\left\{\mu_1(X)-\mu_0(X)-\tau_{\text{ep}}\right\}^2\right]\\ &+\rho\times\text{plim}\left[n_1^{-1}\sum_{j\in S_1}\frac{1-\pi_j}{\pi_j}(1+M^{-1}K_{1,X,j})^2\sigma^2_{A_j}(X_j)\right].\end{aligned}$$

We can construct variance estimators based on the formulas in Theorem 5. However, this again involves estimating the conditional variances $\sigma_0^2(x)$ and $\sigma_1^2(x)$. We recommend using the bootstrap variance estimator in the next subsection.

7.3 A Bootstrap Variance Estimation Procedure

The asymptotic linear forms (23), (25), (27), and (28) are useful for the bootstrap variance estimation. For $b = 1,...,B$, we construct the bootstrap replicates as follows:

Step 1. Sample n_1 units from $\mathcal{S}_1$ with replacement as $\mathcal{S}_1^{*(b)}$.

Step 2. Compute the bootstrap replicates of $\hat{\tau}_2-\tau$ and $\hat{\tau}_{d,\text{ep}}-\tau_{\text{ep}}$ as

$$\begin{aligned}\hat{\tau}_2^{(b)}-\hat{\tau}_2 &= n_1^{-1}\sum_{i\in\mathcal{S}_1^{*(b)}}\pi_i^{-1}I_i\hat{\psi}_i,\\ \hat{\tau}_{1,\text{ep}}^{(b)}-\hat{\tau}_{1,\text{ep}} &= n_1^{-1}\sum_{i\in\mathcal{S}_1^{*(b)}}\hat{\phi}_{1,i},\\ \hat{\tau}_{2,\text{ep}}^{(b)}-\hat{\tau}_{2,\text{ep}} &= n_1^{-1}\sum_{i\in\mathcal{S}_1^{*(b)}}\pi_i^{-1}I_i\hat{\phi}_{2,i},\end{aligned}$$

where $(\hat{\psi}_i,\hat{\phi}_{d,i})$ are the estimated versions of (ψ_i,ϕ_i) from $\mathcal{O}_d$ $d = 1, 2)$.

We estimate Γ, V and v_2 by (16)–(18) based on the above bootstrap replicates, and var $(\hat{\tau})$ by (7), that is, $\hat{v} = \hat{v}_2 - \hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}\hat{\Gamma}$.

Theorem 6.—Under certain regularity conditions, $(\hat{\Gamma}, \hat{V}, \hat{v}_2, \hat{v})$ are consistent for $\{\Gamma, V, \mathrm{var}(\hat{\tau}_2), \mathrm{var}(\hat{\tau})\}$.

For RAL estimators, we can also use the classical nonparametric bootstrap based on resampling the IID observations $\{(I_i, A_i, X_i, U_i, Y_i)\colon\ i \in \mathcal{S}_1\}$ and repeating the analysis as for the original data. The above bootstrap procedure based on resampling the linear forms are particularly useful for the matching estimator.

7.4 Connection With Missing Data

As a final remark, we express the proposed estimator in a linear form that has appeared in the missing data literature.

Proposition 2.—Under certain regularity conditions and Assumption 6, $\hat{\tau}$ has an asymptotic linear form

$$n_1^{1/2}(\hat{\tau} - \tau) = n_1^{-1/2} \sum_{i \in \mathcal{S}_1} \left\{ \frac{I_i}{\pi_i}\psi_i - \left(\frac{I_i}{\pi_i} - 1\right)\Gamma V^{-1}\phi_i \right\} + o_P(1) \tag{29}$$

where ψ_i is $\psi(A_i, X_i, U_i, Y_i)$ for RAL estimators and $\psi_{\mathrm{mat},\,i}$ for the matching estimator, and a similar definition applies to ϕ_i. Under Assumption 1, $\boldsymbol{\pi}_i \equiv \boldsymbol{\rho}$.

Expression (29) is within a class of estimators in the missing data literature with the form

$$n_1^{1/2}(\hat{\tau} - \tau) = n_1^{-1/2} \sum_{i \in \mathcal{S}_1} \left\{ \frac{I_i}{\pi_i} s(A_i, X_i, U_i, Y_i) - \left(\frac{I_i}{\pi_i} - 1\right)\kappa(A_i, X_i, Y_i) \right\} + o_P(1), \tag{30}$$

where $\pi = E(I \mid A, X, U, Y)$, $s(A, X, U, Y)$ satisfies $E\{s(A, X, U, Y)\} = 0$, and $s(A, X, U, Y)$ and $\kappa(A, X, Y)$ are square integrable. Given $s(A, X, U, Y)$ the optimal choice of $\kappa(A, X, Y)$ is $\kappa_{\mathrm{opt}}(A, X, Y) = E\left\{s(A, X, U, Y) \mid A, X, Y\right\}$, which minimizes the asymptotic variance of (30) (Robins, Rotnitzky, and Zhao 1994; Wang et al. 2009). However, $K_{\mathrm{opt}}(A, X, Y)$ requires a correct specification of the missing data model $f(U \mid A, X, Y)$. In our approach, instead of specifying the missing data model, we specify the error-prone estimators and utilize an estimator that is consistent for zero to improve the efficiency of the initial estimator. This is more attractive and closer to empirical practice than calculating $K_{\mathrm{opt}}(A, X, Y)$, because practitioners only need to apply their favorite estimators to the main and validation data using existing software. See also Chen and Chen (2000) for a similar discussion in the regression context under Assumption 1.

8. Discussion

Depending on the roles in statistical inference, there are two types of *big data:* one with large-sample sizes and the other with richer covariates. In our discussion, the main observational data have a larger sample size, and the validation observational data have more covariates. Although some counterexamples exist (Pearl 2009, 2010; Ding and Miratrix 2015; Ding, Vanderweele, and Robins 2017), it is more reliable to draw causal inference from the validation data. The proposed strategy is applicable even the number of covariates is high in the validation data. In this case, we can consider $\hat{\tau}_2$ to be the double machine learning estimators (Chernozhukov et al. 2018) that use flexible machine learning methods for estimating regression and propensity score functions while retain the property in (4). Our framework allows for more efficient estimators of the causal effects by further combining information in the main data, without imposing any parametric models for the partially observed covariates. Coupled with the bootstrap, our estimators require only software implementations of standard estimators, and thus are attractive for practitioners who want to combine multiple observational data sources.

The key insight is to leverage an estimator of zero to improve the efficiency of the initial estimator. If a certain feature is *transportable* across datasets (Bareinboim and Pearl 2016), we can construct a consistent estimator of zero. We have shown that if the validation data are simple random samples from the main data, the distribution of (A, X, Y) is transportable from the validation data to the main data. We then construct a consistent estimator of zero by taking the difference of the estimators based on (A, X, Y) from the two datasets. In the presence of heterogeneity between two data sources, the transportability of the whole distribution of (A, X, Y) can be stringent. However, if we are willing to assume the conditional distribution of Y given (A, X) is transportable, we can then take the error prone estimators to be the regression coefficients of Y on (A, X) from the two datasets. As suggested by one of the reviewers, if the subgroups of two samples are comparable, we can construct the error prone estimators based on the subgroups. Similarly, this construction of error prone estimators can adapt to different transportability assumptions based on the subject matter knowledge.

In the worst case, the heterogeneity is intrinsic between the two samples, and we cannot construct two error prone estimators with the same probability limit. We can still conduct a sensitivity analysis combining two data. Instead of (4), we assume

$$n_2^{1/2}\begin{pmatrix} \hat{\tau}_2 - \tau \\ \hat{\tau}_{2,\mathrm{ep}} - \hat{\tau}_{1,\mathrm{ep}} - \delta \end{pmatrix} \rightarrow \mathcal{N}\left\{0_{L+1}, \begin{pmatrix} v_2 & \Gamma^{\mathrm{T}} \\ \Gamma & V \end{pmatrix}\right\}, \tag{31}$$

where δ is the sensitivity parameter, quantifying the systematic difference between $\hat{\tau}_{2,\mathrm{ep}}$ and $\hat{\tau}_{1,\mathrm{ep}}$. The adjusted estimator becomes $\hat{\tau}_{\mathrm{adj}}(\delta) = \hat{\tau}_2 - \hat{\Gamma}^{\mathrm{T}}\hat{V}^{-1}(\hat{\tau}_{2,\mathrm{ep}} - \hat{\tau}_{1,ep} - \delta)$. With different values of δ, the estimator $\hat{\tau}_{\mathrm{adj}}(\delta)$ can provide valuable insight on the impact of the heterogeneity of the two data, allowing an investigator to assess the extent to which the heterogeneity may alter causal inferences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor, the associate editor, and four anonymous reviewers for suggestions which improved the article significantly. We are grateful to Professor Yi-Hau Chen for providing the data and offering help and advice in interpreting the data. Drs. Lo-Hua Yuan and Xinran Li offered helpful comments. Dr. Yang is partially supported by the National Science Foundation grant DMS 1811245, National Cancer Institute grant P01 CA142538, and Oak Ridge Associated Universities. Dr. Ding is partially supported by the National Science Foundation grant DMS 1713152.

References

Abadie A, and Imbens GW (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” Econometrica, 74, 235–267. [2,4,6,7]

Abadie A (2008), “On the Failure of the Bootstrap for Matching Estimators,” Econometrica, 76, 1537–1557. [6,7]

Abadie A (2016), “Matching on the Estimated Propensity Score,” Econometrica, 84, 781–807. [2]

Angrist JD, Imbens GW, and Rubin DB (1996), “Identification of Causal Effects Using Instrumental Variables,” Journal of American Statistical Association, 91, 444–455. [1]

Antonelli J, Zigler C, and Dominici F (2017), “Guided Bayesian Imputation to Adjust for Confounding When Combining Heterogeneous Data Sources in Comparative Effectiveness Research,” Biostatistics, 18, 553–568. [1] [PubMed: 28334230]

Bang H, and Robins JM (2005), “Doubly Robust Estimation in Missing Data and Causal Inference Models,” Biometrics, 61, 962–973. [2,3] [PubMed: 16401269]

Bareinboim E, and Pearl J (2016), “Causal Inference and the data-fusion problem,” PNAS, 113,7345–7352. [13] [PubMed: 27382148]

Bickel PJ, Klaassen C, Ritov Y, and Wellner J (1993), Efficient and Adaptive Inference in Semiparametric Models, Baltimore: Johns Hopkins University Press [5]

Boonstra PS, Taylor JM, and Mukherjee B (2012), “Incorporating Auxiliary Information for Improved Prediction in High-dimensional Datasets: An Ensemble of Shrinkage Approaches,” Biostatistics, 14, 259–272. [5] [PubMed: 23087411]

Breslow N, McNeney B, and Wellner JA (2003), “Large Sample Theory for Semiparametric Regression Models With Two-phase, Outcome Dependent Sampling,” Annals of Statistics, 31, 1110–1139. [11]

Cao W, Tsiatis AA, and Davidian M (2009), “Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean With Incomplete Data,” Biometrika, 96, 723–734. [2,3] [PubMed: 20161511]

Chang T, and Kott PS (2008), “Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model,” Biometrika, 95, 555–571. [5]

Chatterjee N, Chen YH, Maas P, and Carroll RJ (2016), “Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources,” Journal of American Statistical Association, 111, 107–117. [1,2,7]

Chen J, and Sitter R (1999), “A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys,” Statistica Sinica, 9 385–406. [2,5]

Chen J, Sitter R, and Wu C (2002), “Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys,” Biometrika, 89, 230–237. [2,5]

Chen SX, Leung DHY, and Qin J (2003), “Information Recovery in a Study With Surrogate Endpoints,” Journal of American Statistical Association, 98, 1052–1062. [2,5]

Chen YH (2002), “Cox Regression in Cohort Studies With Validation Sampling,” Journal of Royal Statistical Society, Series B, 64, 51–62. [5]

Chen YH, and Chen H (2000), “A Unified Approach to Regression Analysis Under Double-sampling Designs,” Journal of Royal Statistical Society, Series B, 62, 449–460. [5,13]

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, and Robins J (2018), “Double/debiased Machine Learning for Treatment and Structural Parameters,” The Econometrics Journal, 21, C1–C68. [13]

Cochran WG (2007), Sampling Techniques (3rd ed), New York: Wiley [1,2,8]

Collaboration FS (2009), “Systematically Missing Confounders in Individual Participant Data Meta-analysis of Observational Cohort Studies,” Statistics in Medicine, 28, 1218–1237. [5] [PubMed: 19222087]

Crump R, Hotz VJ, Imbens G, and Mitnik O (2006), “Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand,” Technical Report, 330, Cambridge, MA: National Bureau of Economic Research. Available at http://www.nber.org/papers/T0330 [7]

Deville J-C, and Särndal C-E (1992), “Calibration Estimators in Survey Sampling,” Journal of American Statistical Association, 87, 376–382. [5].

Ding P, and Miratrix LW (2015), “To Adjustor Not to Adjust? Sensitivity Analysis of M-bias and Butterfly-bias,” Journal of Causal Inference, 3, 41–57. [13]

Ding P, Vanderweele T, and Robins J (2017), “Instrumental Variables as Bias Amplifiers With General Outcome and Confounding,” Biometrika, 104,291–302. [13] [PubMed: 29033459]

Efron B, and Tibshirani R (1986), “Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy,” Statistical Science, 1, 54–75. [7]

Enders D, Kollhorst B, Engel S, Linder R, and Pigeot I (2018), “Comparison of Multiple Imputation and Two-phase Logistic Regression to Analyse Two-phase Case-control Studies With Rich Phase 1: A Simulation Study,” Journal of Statistical Computation and Simulation, 88, 2201–2214. [1]

Frangakis CE, and Rubin DB (1999), “Addressing Complications of Intention-to-treat Analysis in the Combined Presence of All-or-none Treatment-noncompliance and Subsequent Missing Outcomes,” Biometrika, 86, 365–379. [3]

Freedman DA (2008), “Randomization Does Not Justify Logistic Regression,” Statistical Science, 23, 237–249. [2]

Fuller WA (2009), Sampling Statistics, Hoboken, NJ: Wiley [5]

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, and Davidian M (2011), “Doubly Robust Estimation of Causal Effects,” American Journal of Epidemiology, 173,761–767. [6] [PubMed: 21385832]

Hájek J (1971), Comment on “An Essay on the Logical Foundations of Survey Sampling, Part One,” by D. Basu, in Foundations of Survey Sampling, eds. Godambe VP and Sprott DA, Toronto: Holt, Rinehart, and Winston, p. 236 [3,11]

Hansen BB (2004), “Full Matching in an Observational Study of Coaching for the SAT,” Journal of American Statistical Association, 99, 609–618. [2]

Hansen LP (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” Econometrica, 50, 1029–1054. [5]

Heckman JJ, Ichimura H and Todd PE (1997), “Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme,” Rev. Econ. Stud, 64, 605–654. [2]

Hirano K, Imbens GW, and Ridder G (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” Econometrica, 71, 1161–1189. [2]

Horvitz DG, and Thompson DJ (1952), “A Generalization of Sampling Without Replacement From a Finite Universe,” Journal of American Statistical Association, 47, 663–685. [2]

Imbens GW (2004), “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” The Review of Economics and Statistics, 86, 4–29. [3]

Imbens GW, and Lancaster T (1994), “Combining Micro and Macro Data in Microeconometric Models,” The Review of Economic Studies, 61,655–680. [1,5]

Jin H, and Rubin DB (2008), “Principal Stratification for Causal Inference With Extended Partial Compliance,” Journal of American Statistical Association, 103, 101–111. [3]

Kim JK, Kwon Y and Paik MC (2016), “Calibrated Propensity Score Method for Survey Nonresponse in Cluster Sampling,” Biometrika, 103,461–473. [5] [PubMed: 27279670]

Kleiner A, Talwalkar A, Sarkar P, and Jordan MI (2014), “A Scalable Bootstrap for Massive Data,” Journal of Royal Statistical Society, Series B, 76, 795–816. [7]

Kott PS (2006), “Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors,” Survey Methodology, 32, 133–142. [5]

Li F, Morgan KL, and Zaslavsky AM (2016), “Balancing Covariates Via Propensity Score Weighting,” Journal of American Statistical Association, 113,390–400. [7]

Lin HW, and Chen YH (2014), “Adjustment for Missing Confounders in Studies Based on Observational Databases: 2-stage Calibration Combining Propensity Scores From Primary and Validation Data,” American Journal of Epidemiology, 180,308–317. [1,2,9] [PubMed: 24966224]

Lumley T, Shaw PA, and Dai JY (2011), “Connections Between Survey Calibration Estimators and Semiparametric Models for Incomplete Data,” International Statistical Review, 79, 200–220. [5] [PubMed: 23833390]

Lunceford JK, and Davidian M (2004), “Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study” Statistical Medicine, 23, 2937–2960. [6]

Lunt M, Glynn RJ, Rothman KJ, Avorn J, and Stürmer T (2012), “Propensity Score Calibration in the Absence of Surrogacy,” American Journal of Epidemiology, 175, 1294–1302. [1] [PubMed: 22688682]

McCandless LC, Richardson S, and Best N (2012), “Adjustment for Missing Confounders Using External Validation Data and Propensity Scores,” Journal of American Statistical Association, 107, 40–51. [1]

Newey WK (1990), “Semiparametric Efficiency Bounds,” Journal of Applied Econometrics, 5, 99–135. [5]

Neyman J (1923), Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T., Statistical Science, 5, 465–472. [2]

Neyman J (1938), “Contribution to the Theory of Sampling Human Populations,” Journal of American Statistical Association, 33, 101–116. [1,2]

Otsu T, and Rai Y (2016), “Bootstrap Inference of Matching Estimators for Average Treatment Effects,” Journal of American Statistical Association, 112, 1720–1732. [7]

Pearl J (2009), “Letter to the Editor: Remarks on the Method of Propensity Score,” Statistics in Medicine, 28, 1420–1423. [13]

Pearl J (2010), “On a Class of Bias-amplifying Variables That Endanger Effect Estimates,” in The Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, eds. Grunwald P and Spirtes P, Corvallis, OR: Association for Uncertainty in Artificial Intelligence, pp. 425–432. [13]

Politis DN, Romano JP, and Wolf M (1999), Subsampling, New York: Springer-Verlag [7]

Qin J (2000), “Combining Parametric and Empirical Likelihoods,” Biometrika, 87, 484–490. [2,5]

Robins JM, Rotnitzky A, and Zhao LP (1994), “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed,” Journal of American Statistical Association, 89, 846–866. [2,5,13]

Rosenbaum PR (1989), “Optimal Matching for Observational Studies,” Journal of American Statistical Association, 84, 1024–1032. [2]

Rosenbaum PR (2002), Studies Observational (2nd ed), New York: Springer [3]

Rosenbaum PR, and Rubin DB (1983a), “Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome,” Journal of Royal Statistical Society, Series B, 45, 212–218. [1]

Rosenbaum PR (1983b), “The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, 70, 41–55. [2,3]

Rosenbaum PR (1973), “Matching to Remove Bias in Observational Studies,” Biometrics,29, 159–183. [2]

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., J. Educ. Psychol 66: 688–701. [2]

Rubin DB (1987). Multiple Imputation for Nonresponse in Surveys, New York: Wiley [9]

Rubin DB (2006). Matched Sampling for Causal Effects, New York: Cambridge University Press [2,3]

Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, and Solomon DH (2005), "Adjusting for Unmeasured Confounders in Pharmacoepidemiologic Claims Data Using External Information: The Example of Cox2 Inhibitors and Myocardial Infarction," Epidemiology, 16, 17–24. [1] [PubMed: 15613941]

Shao J, and Tu D (2012), The Jackknife and Bootstrap, New York: Springer [7]

Stuart EA (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," Statistical Science, 25, 1–21. [2] [PubMed: 20871802]

Stürmer T, Schneeweiss S, Avorn J and Glynn RJ (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration, American Journal of Epidemiology, 162, 279–289. [1,9] [PubMed: 15987725]

Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, and Glynn RJ (2007), "Performance of Propensity Score Calibratio-a Simulation Study, American Journal of Epidemiology, 165, 1110–1118. [1] [PubMed: 17395595]

Tsiatis A (2006), Semiparametric Theory and Missing Data, Springer, New York [3]

Wang W, Scharfstein D, Tan Z, and MacKenzie EJ (2009), "Causal Inference in Outcome-Dependent Two-Phase Sampling Designs," Journal of Royal Statistical Society, Series B, 71, 947–969. [1,2,11,13]

Wang X, and Wang Q (2015), "Semiparametric Linear Transformation Model With Differential Measurement Error and Validation Sampling," Journal of Multivariate Analysis, 141,67–80. [5]

Wu C, and Sitter RR (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data," Journal of American Statistical Association, 96, 185–193. [5]

Yang YW, Chen YH, Wang KH, Wang CY, and Lin HW (2011), "Risk of Herpes Zoster Among Patients With Chronic Obstructive Pulmonary Disease: A Population-based Study," Canadian Medical Association. Journal, 183, 275–280. [9]

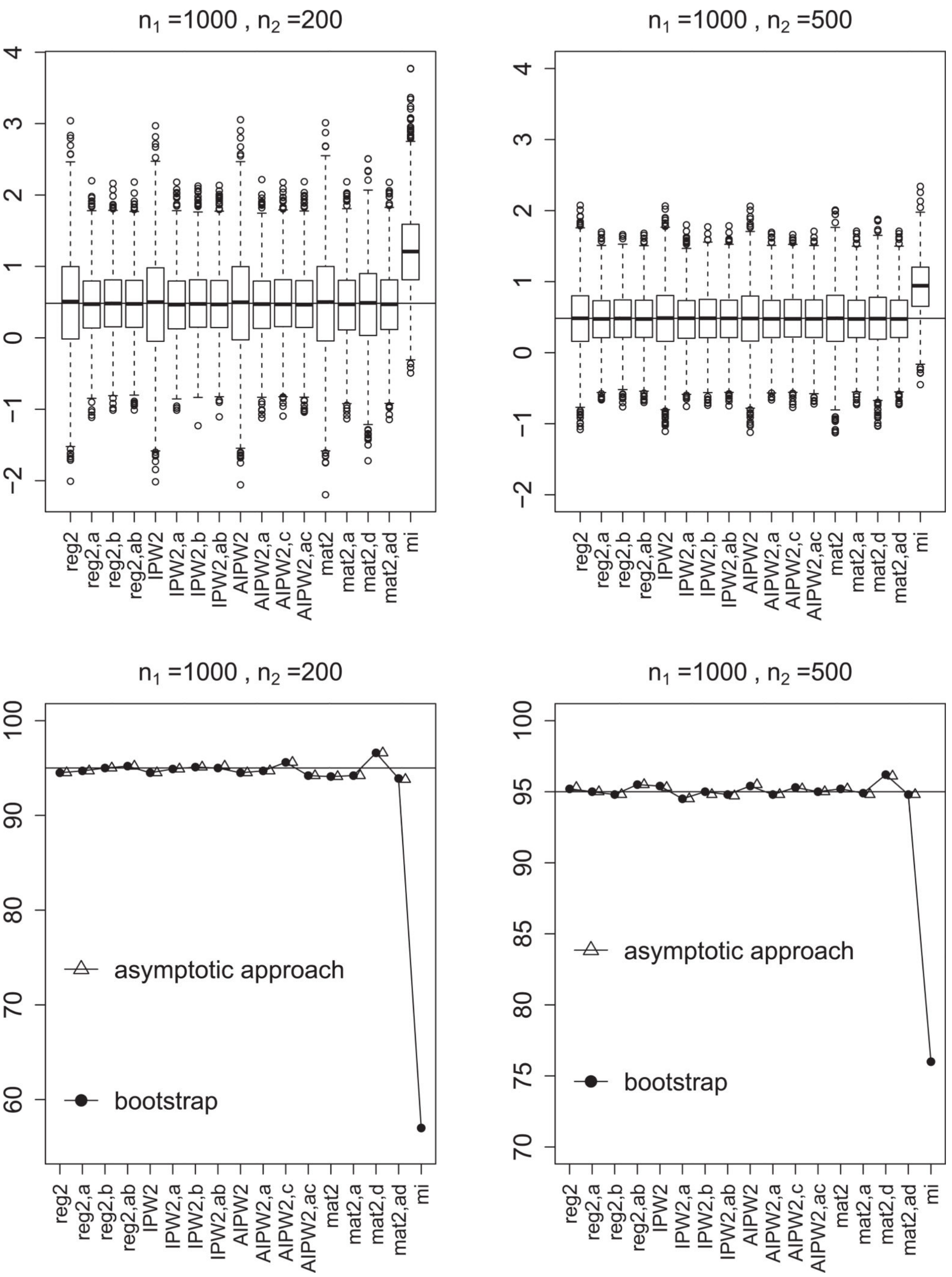

Figure 1.

Simulation results of point estimates (top panels) and coverage rates (bottom panels): the subscripts “a,” “b,” “c,” and “d” stand for methods “reg,” “IPW,” “AIPW,” and “mat,” respectively, “reg2” is $\hat{\tau}_{\text{reg},\,2}$ “reg2,method” is $\hat{\tau}_{\text{reg},\,2\&\text{method}}$, other notation is defined similarly, and “mi” is $\hat{\tau}_{\text{mi}}$

Table 1.

Point estimate, bootstrapped standard error and 95%Wald-type confidence interval

	Est	SE	95% CI		Est	SE	95% CI
$\hat{\tau}_{\text{reg}, 2}$	0.0178	0.0112	(−0.0047, 0.0402)	$\hat{\tau}_{\text{reg}, 2\&\text{reg}}$	0.0155	0.0023	(0.0109, 0.0200)
$\hat{\tau}_{\text{IPW}, 2}$	0.0175	0.0111	(−0.0048, 0.0398)	$\hat{\tau}_{\text{IPW}, 2\&\text{IPW}}$	0.0155	0.0024	(0.0108, 0.0202)
$\hat{\tau}_{\text{AIPW}, 2}$	0.0179	0.0111	(−0.0044, 0.0402)	$\hat{\tau}_{\text{AIPW}, 2\&\text{AIPW}}$	0.0156	0.0024	(0.0109, 0.0203)
$\hat{\tau}_{\text{mat}, 2}$	0.0077	0.0092	(−0.0106, 0.021)	$\hat{\tau}_{\text{mat}, 2\&\text{mat}}$	0.0079	0.0053	(−0.0027, 0.0184)