American Journal of Epidemiology

Vol. 190, No. 6
DOI: 10.1093/aje/kwaa235
Advance Access publication:
March 22, 2021

Practice of Epidemiology

Assessing Heterogeneity of Treatment Effects in Observational Studies

Sarah E. Robertson*, Andrew Leith, Christopher H. Schmid, and Issa J. Dahabreh

* Correspondence to Sarah E. Robertson, Box G-S121-8, Department of Health Services, Policy and Practice, School of Public Health, Brown University, 121 South Main Street, Providence, RI 02912 (e-mail: sarah_robertson@brown.edu).

Initially submitted May 2, 2019; accepted for publication October 16, 2020.

Here we describe methods for assessing heterogeneity of treatment effects over prespecified subgroups in observational studies, using outcome-model–based (g-formula), inverse probability weighting, doubly robust, and matching estimators of subgroup-specific potential outcome means, conditional average treatment effects, and measures of heterogeneity of treatment effects. We compare the finite-sample performance of different estimators in simulation studies where we vary the total sample size, the relative frequency of each subgroup, the magnitude of treatment effect in each subgroup, and the distribution of baseline covariates, for both continuous and binary outcomes. We find that the estimators' bias and variance vary substantially in finite samples, even when there is no unobserved confounding and no model misspecification. As an illustration, we apply the methods to data from the Coronary Artery Surgery Study (August 1975–December 1996) to compare the effect of surgery plus medical therapy with that of medical therapy alone for chronic coronary artery disease in subgroups defined by previous myocardial infarction or left ventricular ejection fraction.

causal inference; effect-measure modification; heterogeneity of treatment effects; observational studies; subgroup analysis

Abbreviations: CASS, Coronary Artery Surgery Study; IPW, inverse probability weighting; OM, outcome model.

Observational studies can be useful for estimating subgroup-specific (conditional) average treatment effects and examining whether these effects are heterogeneous, because, compared with randomized trials, observational studies can have larger sample sizes, leading to more precise effect estimates, especially within subgroups defined by covariates (1, 2). Valid assessment of heterogeneity in observational studies, however, requires the use of statistical methods to control confounding, such as outcome-model (OM)–based (g-formula) (3), inverse probability weighting (IPW) (4, 5), doubly robust (e.g., augmented IPW) (6, 7), or matching methods (8). While the large-sample behavior of these methods is well-understood, their finite-sample performance for assessing heterogeneity of treatment effects is less studied (9–16). In particular, few studies have used simulations to evaluate the finite-sample performance of methods for assessing heterogeneity of treatment effects in observational studies (10–12, 15, 16): 2 studies exclusively evaluated matching estimators (10, 16), 1 study exclusively evaluated regression-based estimators (15), and 2 compared IPW estimators against matching estimators (11, 12). To our knowledge, no study has systematically examined doubly robust estimators for assessing heterogeneity of treatment effects.

In this paper, we consider methods for assessing heterogeneity of treatment effects in observational studies. We focus on the setting where investigators use observational data to assess heterogeneity over prespecified subgroups defined in terms of 1 or more variables selected on the basis of substantive knowledge (e.g., in confirmatory subgroup analysis (1, 2, 17)). In this setting, we describe OM-based, IPW, doubly robust, and matching estimators of subgroup-specific potential outcome means, conditional average treatment effects, and measures of heterogeneity of treatment effects. We compare the performance of different estimators in simulations for continuous and binary outcomes. Lastly, we apply the estimators to data from the Coronary Artery Surgery Study (CASS) to compare the effect of surgery plus medical therapy with the effect of medical therapy alone for chronic coronary artery disease in subgroups defined by previous myocardial infarction and left ventricular ejection fraction.

CAUSAL QUANTITIES

In observational cohort studies investigating heterogeneity of treatment effects for time-fixed (non–time-varying) treatments, the data can be modeled as realizations of independent and identically distributed random tuples (S_i, X_i, A_i, Y_i), $i = 1, \ldots, n$, where S is a baseline (pretreatment) covariate defining the subgroups of interest, X denotes other baseline covariates, A is the treatment, and Y is the observed outcome. For simplicity of exposition, in this paper we consider only binary treatments A and binary candidate effect modifiers S; extensions to multivalued treatments and effect modifiers are straightforward (continuous effect modifiers are outside the scope of this work). Furthermore, we only consider covariates S and X measured at baseline, to avoid conditioning on variables affected by treatment (18, 19). Throughout, we use capital letters for random variables and lowercase letters for realizations of the corresponding random variables.

Let Y_i^a be the potential (counterfactual) outcome for the ith individual under an intervention that sets treatment A to a (20–22). When reporting effect modification findings, a key target of inference is the subgroup-specific potential outcome mean, $E[Y^a|S = s]$, for each treatment a (23). Stratum-specific treatment effect measures are functions of these potential outcome means; for example, the subgroup-specific (conditional) average treatment effect (ATE) is defined as

$$\text{ATE}(S=s) = E[Y^1 - Y^0|S=s] = E[Y^1|S=s] - E[Y^0|S=s].$$

We can quantify heterogeneity over strata of S by comparing stratum-specific treatment effects; for example, we can examine whether $\text{ATE}(S = 1) = \text{ATE}(S = 0)$. The difference between the stratum-specific average treatment effects (dATE), $\text{dATE} = \text{ATE}(S = 1) - \text{ATE}(S = 0)$, quantifies the magnitude of effect modification and is a measure of heterogeneity. A proposal for reporting effect modification findings (23) recommended that investigators report estimates of stratum-specific potential outcome means under the treatments of interest and evaluate treatment effects on both additive and multiplicative scales, when appropriate (e.g., report both the risk difference and the relative risk, for binary outcomes).

IDENTIFICATION

Identifiability conditions

To identify stratum-specific average treatment effects and measures of heterogeneity, we can reexpress their components, the subgroup-specific potential outcome means $E[Y^a|S = s]$, in terms of observed variables. We assume *consistency of potential outcomes:* Among individuals who actually receive treatment a, the potential outcome Y^a equals the observed outcome Y; that is, if $A_i = a$, then $Y_i^a = Y_i$, for every individual i and every treatment a (24). We also assume *"no unmeasured confounding" (conditional exchangeability)* (25), such that the potential outcome under an intervention that sets treatment A to a is independent of treatment conditional on the subgroup indicator and other baseline covariates: $Y^a \perp\!\!\!\perp A|(X, S)$. In observational studies, treatment choice is often influenced by baseline (pretreatment) patient characteristics that are also prognostic factors for the outcome (confounders), inducing dependence between Y^a and A; nevertheless, background knowledge may suggest that the different treatment groups are *exchangeable conditional on covariates*. In practice, this exchangeability assumption is untestable using the observed data, and sensitivity analyses are needed to assess the impact of potential violations on study results (26). Finally, we assume *positivity of the conditional probability of treatment:* If $f_{X,S}(x, s) \neq 0$, then $\Pr[A = a|X = x, S = s] > 0$.

Identification of the subgroup-specific potential outcome mean

As we show in Web Appendix 1 (available online at https://doi.org/10.1093/aje/kwaa235), when the identifiability conditions hold, we can rewrite the subgroup-specific potential outcome mean for treatment a using the observed data as

$$E\left[Y^a|S = s\right] = E\left[E\left[Y|X, S, A = a\right]|S = s\right] \equiv \mu\left(a, s\right). \tag{1}$$

The functional $\mu(a, s)$ can be estimated in observational studies that collect information on (S, X, A, Y) (3).

Under positivity, $\mu(a, s)$ has an algebraically equivalent IPW (27) reexpression,

$$\mu\left(a, s\right) = \frac{1}{\Pr[S = s]} E\left[\frac{I(S = s, A = a)\,Y}{\Pr[A = a|X, S]}\right], \tag{2}$$

where $I(\cdot)$ denotes the indicator function. Furthermore, under positivity,

$$\mu\left(a, s\right) = \left\{E\left[\frac{I(S = s, A = a)}{\Pr[A = a|X, S]}\right]\right\}^{-1} E\left[\frac{I(S = s, A = a)\,Y}{\Pr[A = a|X, S]}\right]. \tag{3}$$

It follows that subgroup-specific average treatment effects can be identified by taking differences and that measures of heterogeneity can be identified as contrasts of these subgroup-specific effects. For example, $\text{ATE}(S = s) = \mu(1, s) - \mu(0, s)$ and $\text{dATE} = \mu(1, 1) - \mu(0, 1) - \mu(1, 0) + \mu(0, 0)$.

ESTIMATION AND INFERENCE

We now describe methods for estimation and inference for subgroup-specific potential outcome means, which are the components of subgroup-specific average treatment effects and measures of heterogeneity. Throughout, we assume that parametric models are used to estimate conditional probabilities or expectations, because this is the most common approach in practice. In the Discussion section, we consider the use of data-adaptive modeling methods (e.g., machine learning).

Estimation

Outcome modeling and standardization. The identification result in equation 1 suggests the following OM-based (g-formula) estimator for the potential outcome mean under treatment a for stratum s,

$$\hat{\mu}_{\mathrm{OM}}(a,s) = \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} I(S_i = s)\, \hat{g}_a(X_i, S_i), \tag{4}$$

where $\hat{g}_a(X, S)$ is an estimator for $E[Y|X, S, A = a]$. Typically, we use a parametric model for the conditional outcome mean in each treatment group, $g_a(X, S; \theta_a)$, with finite dimensional parameter θ_a. Separate models for the relationship of Y with X can be estimated in each treatment group and each stratum of S, if the data allow it. In many applications, however, because of the curse of dimensionality, it will be necessary to "borrow strength" by assuming some degree of homogeneity across strata of S (even if separate models are fitted within each treatment group). When the parametric outcome model $g_a(X, S; \theta_a)$ is correctly specified, the estimator $\hat{\mu}_{\mathrm{OM}}(a, s)$ converges in probability to $\mu(a, s)$ (25).

Inverse probability weighting. The IPW reexpressions of the identification results suggest 2 IPW estimators: one based on equation 2,

$$\hat{\mu}_{\mathrm{IPW1}}(a,s) = \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i)\, Y_i, \tag{5}$$

and the other based on equation 3,

$$\hat{\mu}_{\mathrm{IPW2}}(a,s) = \left\{ \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i) \right\}^{-1} \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i)\, Y_i, \tag{6}$$

where $\hat{w}_{a,s}(X_i, S_i, A_i) = \frac{I(S_i = s, A_i = a)}{\hat{e}_a(X_i, S_i)}$ and $\hat{e}_a(X, S)$ is an estimator for $\Pr[A = a|X, S]$ (i.e., the propensity score (4, 5)).

When the treatment A is binary, estimates of $\Pr[A = a|X, S]$ are typically obtained using a logistic regression model, $e_a(X, S; \gamma)$, with finite dimensional parameter γ. Separate models for the relationship of A with X can be estimated within each stratum of S, if the data allow it. When the parametric model for the probability of treatment $e_a(X, S; \gamma)$ is correctly specified, both estimators, $\hat{\mu}_{\mathrm{IPW1}}(a, s)$ and $\hat{\mu}_{\mathrm{IPW2}}(a, s)$, converge in probability to $\mu(a, s)$. When weights are highly variable, $\hat{\mu}_{\mathrm{IPW2}}(a, s)$ should usually be preferred over $\hat{\mu}_{\mathrm{IPW1}}(a, s)$ because it normalizes the weights by their sum (28) and produces estimates that always fall within the support of the outcome variable (29). For example, estimates of the potential outcome mean for a binary outcome are always between 0 and 1 from $\hat{\mu}_{\mathrm{IPW2}}(a, s)$; in contrast, estimates from $\hat{\mu}_{\mathrm{IPW1}}(a, s)$ can be greater than 1.

Doubly robust estimators. For consistent estimation, the OM-based estimator and the IPW estimator rely on correct specification of models for the expectation of the outcome or the probability of treatment, respectively. To gain some robustness to model misspecification, it may be advantageous to use estimators that combine both models and are doubly robust (7), in the sense that they are consistent when either model is correctly specified. When both models are correctly specified and converge at a sufficiently fast rate, doubly robust estimators of the stratum-specific outcome mean are semiparametric efficient (30–32). Perhaps more important for practical applications, doubly robust estimators often produce estimates that are more precise than those produced by IPW estimators (even when the outcome model is not correctly specified) (33). One doubly robust (DR) estimator is

$$\begin{aligned}\hat{\mu}_{\mathrm{DR1}}(a,s) &= \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i) \left\{ Y_i - \hat{g}_a(X_i, S_i) \right\} \\ &\quad + \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} I(S_i = s)\, \hat{g}_a(X_i, S_i),\end{aligned} \tag{7}$$

with $\hat{w}_{a,s}(X, S, A)$ and $\hat{g}_a(X, S)$ as defined above.

Normalizing the weights, we obtain a second doubly robust estimator:

$$\begin{aligned}\hat{\mu}_{\mathrm{DR2}}(a,s) &= \left\{ \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i) \right\}^{-1} \sum_{i=1}^{n} \hat{w}_{a,s}(X_i, S_i, A_i) \left\{ Y_i - \hat{g}_a(X_i, S_i) \right\} \\ &\quad + \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} I(S_i = s)\, \hat{g}_a(X_i, S_i).\end{aligned} \tag{8}$$

A third doubly robust estimator relies on fitting a multivariable regression model for the outcome estimated using IPW (with weights as described above), followed by standardization over the distribution of baseline covariates (29, 34). The potential outcome mean is estimated as

$$\hat{\mu}_{\mathrm{DR3}}(a,s) = \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} I(S_i = s)\, g_a(X_i, S_i; \widetilde{\theta}), \tag{9}$$

where $g_a(X, S; \widetilde{\theta})$ is an estimator for $E[Y|X, S, A = a]$ obtained by a multivariable outcome model estimated by IPW. When the outcome is modeled in the linear exponential family with a canonical link, and estimation is carried out using quasilikelihood methods (e.g., linear or logistic regression) (35), this estimator has the double robustness property.

Notably, $\hat{\mu}_{\mathrm{DR3}}(a, s)$ always produces estimates that fall within the support of the distribution of the outcome (as long as the outcome model is reasonable); while neither $\hat{\mu}_{\mathrm{DR1}}(a, s)$ nor $\hat{\mu}_{\mathrm{DR2}}(a, s)$ has this property, $\hat{\mu}_{\mathrm{DR2}}(a, s)$ uses normalized weights that sum to the subgroup sample size and often is more well-behaved than $\hat{\mu}_{\mathrm{DR1}}(a, s)$ in finite samples. Asymptotically, all of the estimators converge in probability to $\mu(a, s)$ when either the model for the probability of treatment or the model for the mean of the outcome is correctly specified (29, 34).

Matching. Matching is probably the most popular approach for estimating causal effects in applied work (e.g., in the literature on ischemic heart disease (36, 37), from which our data example stems), including in work assessing heterogeneity of treatment effects using observational data (11, 16, 38). For comparison with the methods described above, we implemented a simple matching estimator that relied on nearest-neighbor 1:1 Mahalanobis distance matching on the estimated propensity score, with replacement (39).

This matching estimator can be thought of as an imputation estimator for individual-level potential outcomes (40). Specifically, for unit i in subgroup s, let $j(i)$ be the index $j \in \{1, \ldots, n\}$ that satisfies $A_j = 1 - A_i$, $S_j = S_i$ and selects the nearest neighbor of unit i in subgroup s. Then, for 1:1 matching, the matching estimator imputes individual-level potential outcomes, Y_i^a under treatment a, as

$$\tilde{Y}_i^a = \begin{cases} Y_i & \text{if } A_i = a; \\ Y_{j(i)} & \text{if } A_i \neq a. \end{cases}$$

The matching (MT) estimator of $\mu(a, s)$ is

$$\hat{\mu}_{\mathrm{MT}}(a, s) = \left\{ \sum_{i=1}^{n} I(S_i = s) \right\}^{-1} \sum_{i=1}^{n} I(S_i = s)\, \tilde{Y}_i^a. \quad (10)$$

In large samples, this estimator converges in probability to $\mu(a, s)$, provided that the propensity score model is correctly specified (40, 41) (large-sample properties for matching estimators are not as straightforward to derive as for other estimators in this paper; here, we focus on a fairly simple estimator).

Inference

For all of the estimators described above except matching, inference using standard M-estimation methods (42) is straightforward when using parametric working models (e.g., see Lunceford and Davidian (43) and Williamson et al. (44)). Inference based on the nonparametric bootstrap (45) is also easy to obtain and will often be preferred in practice. Inference about matching estimators, especially those that are more complicated than the simple estimator we use in this paper, is more challenging because of the nonsmooth nature of matching procedures; however, both large-sample approximations and specialized bootstrap procedures are available (46, 47).

SIMULATION STUDY

We conducted a simulation study to compare the finite-sample behavior of OM-based, IPW, doubly robust, and matching estimators in observational studies with no unmeasured confounding and correctly specified parametric models. Though our simulation reflects a "best-case scenario" in which the causal *and* statistical modeling assumptions are satisfied, it suffices to uncover important differences among the estimators we consider. We examined both continuous and binary outcomes across different scenarios where we varied the sample size, the subgroup prevalence, the magnitude of the treatment effect, and the correlation between baseline covariates. Here in the main text of this article, we describe the simulation methods and report selected results for continuous outcomes; in Web Appendix 2, we provide additional implementation details for binary outcomes. Software code with which to replicate our simulations is available on GitHub (see Web Appendix 3 for the link to our GitHub repository).

Targets of inference and measures of performance

For each simulation scenario, we generated data, applied the estimators, and used 100,000 replications to assess performance. Specifically, we estimated the bias and standard deviation when using each of the estimators described in the previous section for subgroup-specific potential outcome means and for differences between subgroup-specific average treatment effects that quantify effect modification. We scaled (multiplied) bias and standard deviation results by $\sqrt{n}$ to facilitate comparisons of the behavior of different estimators as n increases; if $\hat{\mu}(a, s)$ is an asymptotically normal estimator of $\mu(a, s)$, the $\sqrt{n}$-scaled difference between $\hat{\mu}(a, s)$ and $\mu(a, s)$—that is, $\sqrt{n}(\hat{\mu}(a, s) - \mu(a, s))$—converges to a mean-zero normal distribution as $n \to \infty$ (25, 48).

Simulation study setup

We use $X = (1, X_1, \ldots, X_8)^{\mathrm{T}}$ to denote the vector of pretreatment covariates other than the effect modifier under consideration; the covariates in X are of 2 kinds: *confounding variables*, $X_1, \ldots, X_4$, which affect the treatment and the outcome, and *"pure" outcome predictors*, $X_5, \ldots, X_8$, which affect the outcome but not the treatment (after conditioning on $X_1, \ldots, X_4$). As above, S denotes a binary baseline (pretreatment) covariate for the subgroups of interest, A the treatment received, and Y the observed outcome.

Data generation for *S*, *X*, and *A*

We generated the subgroup indicator S as a Bernoulli random variable with parameter $\Pr[S = 1]$. We considered both balanced (equal prevalence) subgroups with $\Pr[S = 1] = 0.5$ and imbalanced (unequal prevalence) subgroups with $\Pr[S = 1] = 0.25$. We generated all other baseline covariates from independent standard normal distributions, $X_j \sim \mathcal{N}(0, 1), j = 1, \ldots, 8$. We also considered a case where the baseline covariates in X may be correlated with each other, following a standard normal distribution with a correlation varying between −0.2 and 0.2 (the full correlation structure is provided in Web Appendix 2). We simulated the treatment choice in each subgroup using a binary indicator

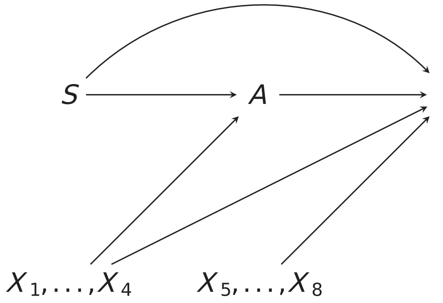

Figure 1. Structural model used to simulate the performance of different estimators of heterogeneity of treatment effects with independent baseline covariates. S is the subgroup indicator, $X_j, j = 1, \ldots, 8$ are additional baseline covariates, A is the treatment variable, and Y is the observed outcome. Confounding variables, $X_j, j = 1, \ldots, 4$, affect both A and Y, while pure outcome predictors, $X_j, j = 5, \ldots, 8$, affect only Y. Note that S is structurally a (potential) confounder; we graph it separately to emphasize that it is the variable over which heterogeneity is to be examined.

A with $\Pr[A = 1|S = s, X] = \text{expit}(\alpha_s X)$ and $\alpha_s = (0,\ \phi_s,\ \ln(3),\ \ln(2), \ln(3),\ 0,\ 0,\ 0,\ 0)$, so the value of the coefficient of X_1 in $S = 1$, $\phi_1 = \ln(1.5)$ corresponded to a conditional odds ratio that was half the value in $S = 0$, $\phi_0 = \ln(3)$, but all other coefficients of baseline covariates were the same.

Figure 1 is a directed acyclic graph showing the data-generating mechanism in the simulations where the covariates in X are independently drawn. Web Appendix 2 contains Web Figure 1, the corresponding graph for the simulation with correlated covariates.

Data generation for continuous *Y*

We generated the observed continuous outcome, for subgroup $S = s$ and treatment group $A = a$, as $Y = \beta_{a,s}X + \epsilon$, with $\beta_{a,s} = (\lambda_{a,s}, \zeta_s, 1.5, 1.5, 1, 1, 1, 1, 1)$, where $\lambda_{a,s}$ depends on subgroup and treatment and the coefficient of X_1 depends on subgroup, with $\zeta_1 = 3$ and $\zeta_0 = 1.5$. Coefficients of $\lambda_{a,s}$ were chosen to obtain a desired subgroup-specific average treatment effect; ϵ had an independent standard normal distribution. We examined 4 different average treatment effects for $S = 1$: namely, $\text{ATE}(S = 1) = 0$, $-0.5, -1$, or -2. For each of these treatment effects, we examined 4 different versions of effect modification: $\text{ATE}(S = 0) = \text{ATE}(S = 1)$ (absence of effect modification); $\text{ATE}(S = 0) = -0.5 + \text{ATE}(S = 1)$; $\text{ATE}(S = 0) = -1 + \text{ATE}(S = 1)$; and $\text{ATE}(S = 0) = -\text{ATE}(S = 1)$ (qualitative effect modification, when the treatment effect is not null). Web Table 1 in Web Appendix 2 lists all of the parameter values for the simulation study for continuous outcomes.

Model specification in the simulation study

We used correctly specified parametric models. We modeled the probability of treatment using a logistic regression model with main effects for all pretreatment covariates, including confounding variables and pure outcome predictors, $X_1, \ldots, X_8$ and S, and the product term $X_1 \times S$. We also modeled the probability of treatment using a logistic regression model that included only confounding variables, $X_1, \ldots, X_4$ and S, and the product term $X_1 \times S$. We always modeled the expectation of the continuous outcome Y using a linear regression model with main effects for $X_1, \ldots, X_8$ and S, the product term $X_1 \times S$, treatment A, and the product term $A \times S$.

Software

We implemented the numerical methods described in Web Appendix 2 to determine simulation parameters for binary outcomes in R, version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria). All simulations were carried out in Stata, version MP/15.1 (StataCorp LLC, College Station, Texas).

SIMULATION RESULTS

Here in the main text, we report simulation results from scenarios with continuous outcomes, $\Pr[S = 1] = 0.5$, and independent baseline covariates, when estimators rely on models that include all confounders and pure outcome predictors. We briefly summarize results for other continuous outcome simulation scenarios and for binary outcomes; we report detailed results from these simulations in Web Appendix 2. Web Table 2 in Web Appendix 2 lists the full parameter values for the simulation study for binary outcomes. Furthermore, in this paper, we report results for analyses in which the target parameter is the difference between the stratum-specific average treatment effects. We use this "difference-in-differences" parameter as a convenient summary of the simulation results (results for subgroup-specific potential outcome means are available on GitHub; see Web Appendix 3 for a link).

Bias

Table 1 shows that the OM-based estimator and the 3 doubly robust estimators had little $\sqrt{n}$-scaled bias, even in small sample sizes. In contrast, IPW estimators and the matching estimator had substantial bias with small sample sizes. This bias became smaller for IPW estimators as the sample size increased, and for the matching estimator (in exploratory simulations with a sample size of 100,000), but the decrease was most pronounced for the IPW estimator without normalized weights. In general, the IPW estimator with normalized weights had larger bias than the IPW estimator with nonnormalized weights for all sample sizes.

Standard deviation

Table 2 shows that OM-based estimators had the lowest $\sqrt{n}$-scaled standard deviation; doubly robust estimators had a standard deviation slightly larger than that of OM-based

Table 1. Bias Multiplied by $\sqrt{n}$ for Estimators of the Difference Between Average Treatment Effects in a Simulation Study[a]

Sample Size (n)	ATE[b]		dATE[c]	Estimator[d]						
	ATE(S = 1)	ATE(S = 0)		OM	IPW1	IPW2	DR1	DR2	DR3	MT
500[e]	0.0	0.0	0.0	−0.005	−0.296	−1.618	0.014	0.006	−0.002	−0.968
500	0.0	−0.5	0.5	0.008	0.031	−1.487	0.021	0.015	0.026	−0.917
500	0.0	−1.0	1.0	−0.013	−0.191	−1.463	0.001	−0.008	−0.012	−0.934
500[e]	0.0	0.0	0.0	0.009	−0.101	−1.501	0.014	0.011	0.007	−1.043
500	−0.5	−0.5	0.0	0.012	0.280	−1.259	0.026	0.008	0.008	−0.829
500	−0.5	−1.0	0.5	−0.022	−0.142	−1.431	−0.024	−0.018	−0.016	−0.880
500	−0.5	−1.5	1.0	0.015	0.213	−1.275	0.008	−0.002	−0.001	−0.856
500	−0.5	0.5	−1.0	−0.015	0.081	−1.507	0.015	−0.006	−0.009	−0.957
500	−1.0	−1.0	0.0	−0.005	−0.328	−1.618	0.014	0.006	−0.002	−0.968
500	−1.0	−1.5	0.5	0.008	0.013	−1.487	0.021	0.015	0.026	−0.917
500	−1.0	−2.0	1.0	−0.013	−0.207	−1.463	0.001	−0.008	−0.012	−0.934
500	−1.0	1.0	−2.0	0.009	−0.085	−1.501	0.014	0.011	0.007	−1.043
500	−2.0	−2.0	0.0	0.012	0.296	−1.259	0.026	0.008	0.008	−0.829
500	−2.0	−2.5	0.5	−0.022	−0.138	−1.431	−0.024	−0.018	−0.016	−0.880
500	−2.0	−3.0	1.0	0.015	0.241	−1.275	0.008	−0.002	−0.001	−0.856
500	−2.0	2.0	−4.0	−0.015	0.112	−1.507	0.015	−0.006	−0.009	−0.957
1,000[e]	0.0	0.0	0.0	−0.000	−0.066	−1.318	−0.004	−0.009	−0.015	−0.906
1,000	0.0	−0.5	0.5	0.019	−0.038	−1.354	0.028	0.022	0.020	−1.069
1,000	0.0	−1.0	1.0	−0.007	−0.033	−1.308	−0.010	−0.004	−0.005	−1.013
1,000[e]	0.0	0.0	0.0	0.007	−0.130	−1.303	0.018	0.011	0.007	−0.927
1,000	−0.5	−0.5	0.0	0.030	0.077	−1.244	0.023	0.034	0.040	−0.961
1,000	−0.5	−1.0	0.5	−0.005	−0.188	−1.380	0.019	0.010	0.004	−0.967
1,000	−0.5	−1.5	1.0	−0.016	−0.166	−1.446	0.000	−0.004	−0.008	−1.007
1,000	−0.5	0.5	−1.0	−0.008	−0.076	−1.333	−0.002	−0.004	−0.006	−0.978
1,000	−1.0	−1.0	0.0	−0.000	−0.085	−1.318	−0.004	−0.009	−0.015	−0.906
1,000	−1.0	−1.5	0.5	0.019	−0.041	−1.354	0.028	0.022	0.020	−1.069
1,000	−1.0	−2.0	1.0	−0.007	−0.038	−1.308	−0.010	−0.004	−0.005	−1.013
1,000	−1.0	1.0	−2.0	0.007	−0.125	−1.303	0.018	0.011	0.007	−0.927
1,000	−2.0	−2.0	0.0	0.030	0.088	−1.244	0.023	0.034	0.040	−0.961
1,000	−2.0	−2.5	0.5	−0.005	−0.221	−1.380	0.019	0.010	0.004	−0.967
1,000	−2.0	−3.0	1.0	−0.016	−0.188	−1.446	0.000	−0.004	−0.008	−1.007
1,000	−2.0	2.0	−4.0	−0.008	−0.034	−1.333	−0.002	−0.004	−0.006	−0.978
5,000[e]	0.0	0.0	0.0	0.021	0.121	−0.738	0.046	0.042	0.036	−1.030
5,000	0.0	−0.5	0.5	−0.011	0.100	−0.723	−0.027	−0.027	−0.023	−1.086
5,000	0.0	−1.0	1.0	−0.002	0.215	−0.671	0.018	0.017	0.015	−0.995
5,000[e]	0.0	0.0	0.0	−0.003	0.316	−0.531	0.019	0.021	0.017	−0.960
5,000	−0.5	−0.5	0.0	0.011	−0.159	−0.927	0.013	0.016	0.022	−1.022
5,000	−0.5	−1.0	0.5	0.025	0.105	−0.709	0.001	0.000	0.001	−1.111
5,000	−0.5	−1.5	1.0	−0.010	−0.065	−0.855	−0.005	−0.011	−0.016	−1.040
5,000	−0.5	0.5	−1.0	0.005	0.004	−0.845	0.008	0.006	0.008	−1.053
5,000	−1.0	−1.0	0.0	0.021	0.112	−0.738	0.046	0.042	0.036	−1.030
5,000	−1.0	−1.5	0.5	−0.011	0.108	−0.723	−0.027	−0.027	−0.023	−1.086

Table continues

Table 1. Continued

Sample Size (n)	ATE[b]		dATE[c]	Estimator[d]						
	ATE(S = 1)	ATE(S = 0)		OM	IPW1	IPW2	DR1	DR2	DR3	MT
5,000	−1.0	−2.0	1.0	−0.002	0.219	−0.671	0.018	0.017	0.015	−0.994
5,000	−1.0	1.0	−2.0	−0.003	0.333	−0.531	0.019	0.021	0.017	−0.960
5,000	−2.0	−2.0	0.0	0.011	−0.172	−0.927	0.013	0.016	0.022	−1.022
5,000	−2.0	−2.5	0.5	0.025	0.139	−0.709	0.001	0.000	0.001	−1.111
5,000	−2.0	−3.0	1.0	−0.010	−0.073	−0.855	−0.005	−0.011	−0.016	−1.040
5,000	−2.0	2.0	−4.0	0.005	0.009	−0.845	0.008	0.006	0.008	−1.053

Abbreviations: ATE, average treatment effect; dATE, difference between the stratum-specific average treatment effects; DR, doubly robust; IPW, inverse probability weighting; MT, matching; OM, outcome modeling.

[a] Results for continuous outcome simulations with Pr[$S = 1$] = 0.5 and independent baseline covariates.

[b] ATE($S = s$), mean difference in stratum s.

[c] ATE($S = 1$) − ATE($S = 0$).

[d] Notation: OM, OM estimator in equation 4; IPW1, IPW estimator in equation 5; IPW2, IPW estimator in equation 6; DR1, DR estimator in equation 7; DR2, DR estimator in equation 8; DR3, DR estimator in equation 9; MT, MT estimator in equation 10.

[e] For each set of results by sample size (n), the fourth row has the same simulation parameters as the first row because setting ATE($S = 1$) = 0 leads to ATE($S = 0$) = 0, both when ATE($S = 0$) = ATE($S = 1$) and when ATE($S = 0$) = −ATE($S = 1$). Results may differ between the first and fourth rows due to simulation error.

estimators but much lower than those of matching and IPW estimators. The IPW estimator with normalized weights had a smaller standard deviation than the IPW estimator with nonnormalized weights.

Additional simulation results for continuous outcomes

Web Tables 3–8 in Web Appendix 4 show the results for simulation scenarios with correlated baseline covariates and with Pr[$S = 1$] = 0.25. We found similar trends and magnitudes in terms of bias and standard deviation, as described above. When subgroups were imbalanced in terms of their prevalence (Pr[$S = 1$] = 0.25) we found that estimators generally had more bias and higher variance when compared with the simulation scenarios with balanced subgroup prevalences (Pr[$S = 1$] = 0.5).

Simulation results for binary outcomes

Web Tables 9–20 in Web Appendix 5 show results for binary outcome simulations using 10,000 runs for each scenario (the smaller number of runs compared with continuous outcomes was deemed appropriate because the simulation error was smaller for binary outcomes). Trends in the estimators' bias and variance were similar to the results from simulations with continuous outcomes. Occasionally, convergence issues (caused by extreme weights) prevented estimation using the estimator in equation 9 (DR3); this occurred in fewer than 1% of all simulation runs and in only 1 simulation scenario (scenario 16), which has relatively large subgroup-specific potential outcome means. Notably, for binary outcomes, some of the estimators can return potential outcome mean estimates that are lower than 0 or greater than 1 (this can happen for estimators that can produce estimates that do not fall within the support of the distribution of the outcome Y—that is, the IPW estimator with nonnormalized weights and the 2 doubly robust estimators that are not based on weighted regression.

ILLUSTRATION OF THE METHODS IN CASS

CASS design and data

To illustrate the methods, we used data (August 1975–December 1996) from CASS, a comprehensive cohort study (49) that compared coronary artery bypass grafting surgery plus medical therapy (henceforth, "surgery") with medical therapy alone for patients with chronic coronary artery disease. Details about the design of CASS are available elsewhere (50, 51). In brief, patients undergoing angiography in 11 institutions were screened for eligibility. From a total of 2,099 eligible patients, 780 consented to randomization and 1,319 declined and were enrolled in an observational study of the same treatments. We excluded 6 patients for consistency with prior CASS analyses (52, 53) and in accordance with CASS data release notes.

We separately analyzed individuals in the observational and randomized components of CASS to estimate the 10-year mortality risk (cumulative incidence proportion), risk difference, and risk ratio among subgroups defined by previous myocardial infarction (about 60% of patients) or ejection fraction ≥50% (about 80%). No patients were lost to follow-up in the first 10 years of the study; therefore, cumulative incidence proportions are reasonable measures of incidence. For individuals in the trial, because randomization ensures exchangeability, we used an unadjusted analysis. For individuals in the observational study, we used outcome regression

Table 2. Standard Deviation Multiplied by $\sqrt{n}$ for Estimators of the Difference Between Average Treatment Effects in a Simulation Study[a]

Sample Size (n)	ATE[b]		dATE[c]	Estimator[d]						
	ATE(S = 1)	ATE(S = 0)		OM	IPW1	IPW2	DR1	DR2	DR3	MT
500[e]	0.0	0.0	0.0	4.180	53.224	33.413	8.310	6.689	5.961	19.890
500	0.0	−0.5	0.5	4.198	63.124	33.276	9.386	6.706	5.973	19.933
500	0.0	−1.0	1.0	4.207	56.154	33.273	8.395	6.715	5.979	20.030
500[e]	0.0	0.0	0.0	4.191	51.647	33.123	8.257	6.694	5.964	19.987
500	−0.5	−0.5	0.0	4.194	53.923	33.252	8.250	6.716	5.972	19.939
500	−0.5	−1.0	0.5	4.186	55.147	33.233	8.122	6.695	5.971	20.036
500	−0.5	−1.5	1.0	4.193	57.115	33.402	8.293	6.718	5.991	20.002
500	−0.5	0.5	−1.0	4.197	53.988	33.124	8.895	6.710	5.996	19.990
500	−1.0	−1.0	0.0	4.180	56.423	33.413	8.310	6.689	5.961	19.890
500	−1.0	−1.5	0.5	4.198	65.909	33.276	9.386	6.706	5.973	19.933
500	−1.0	−2.0	1.0	4.207	59.634	33.273	8.395	6.715	5.979	20.030
500	−1.0	1.0	−2.0	4.191	50.782	33.123	8.257	6.694	5.964	19.987
500	−2.0	−2.0	0.0	4.194	58.845	33.252	8.250	6.716	5.972	19.939
500	−2.0	−2.5	0.5	4.186	59.968	33.233	8.122	6.695	5.971	20.036
500	−2.0	−3.0	1.0	4.193	62.706	33.402	8.293	6.718	5.991	20.002
500	−2.0	2.0	−4.0	4.197	52.605	33.124	8.895	6.710	5.996	19.990
1,000[e]	0.0	0.0	0.0	4.181	47.983	36.048	7.961	6.931	6.262	20.517
1,000	0.0	−0.5	0.5	4.154	51.191	36.320	8.102	6.948	6.254	20.467
1,000	0.0	−1.0	1.0	4.160	58.206	35.916	8.392	6.918	6.255	20.532
1,000[e]	0.0	0.0	0.0	4.164	48.328	36.109	7.882	6.953	6.273	20.542
1,000	−0.5	−0.5	0.0	4.164	56.436	36.093	7.964	6.946	6.259	20.584
1,000	−0.5	−1.0	0.5	4.173	50.983	36.181	7.837	6.943	6.273	20.628
1,000	−0.5	−1.5	1.0	4.148	53.126	36.182	7.914	6.915	6.238	20.617
1,000	−0.5	0.5	−1.0	4.163	46.729	35.842	7.901	6.925	6.255	20.527
1,000	−1.0	−1.0	0.0	4.181	50.895	36.048	7.961	6.931	6.262	20.517
1,000	−1.0	−1.5	0.5	4.154	54.325	36.320	8.102	6.948	6.254	20.467
1,000	−1.0	−2.0	1.0	4.160	60.764	35.916	8.392	6.918	6.255	20.532
1,000	−1.0	1.0	−2.0	4.164	47.927	36.109	7.882	6.953	6.273	20.542
1,000	−2.0	−2.0	0.0	4.164	61.549	36.093	7.964	6.946	6.259	20.584
1,000	−2.0	−2.5	0.5	4.173	55.683	36.181	7.837	6.943	6.273	20.628
1,000	−2.0	−3.0	1.0	4.148	58.284	36.182	7.914	6.915	6.238	20.617
1,000	−2.0	2.0	−4.0	4.163	46.110	35.842	7.901	6.925	6.255	20.527
5,000[e]	0.0	0.0	0.0	4.140	47.303	41.677	7.654	7.304	6.833	21.782
5,000	0.0	−0.5	0.5	4.150	47.486	41.390	7.698	7.320	6.844	21.829
5,000	0.0	−1.0	1.0	4.161	53.405	41.765	7.728	7.321	6.856	21.866
5,000[e]	0.0	0.0	0.0	4.152	47.555	41.394	7.793	7.291	6.827	21.783
5,000	−0.5	−0.5	0.0	4.159	46.686	41.189	7.548	7.280	6.834	21.843
5,000	−0.5	−1.0	0.5	4.136	50.979	41.335	7.619	7.291	6.832	21.812
5,000	−0.5	−1.5	1.0	4.153	49.273	41.428	7.666	7.337	6.859	21.817
5,000	−0.5	0.5	−1.0	4.148	47.167	41.634	7.627	7.308	6.852	21.842
5,000	−1.0	−1.0	0.0	4.140	50.059	41.677	7.654	7.304	6.833	21.782
5,000	−1.0	−1.5	0.5	4.150	50.377	41.390	7.698	7.320	6.844	21.830
5,000	−1.0	−2.0	1.0	4.161	56.144	41.765	7.728	7.321	6.856	21.866

Table continues

Table 2. Continued

Sample Size (n)	ATE[b]		dATE[c]	Estimator[d]						
	ATE(S = 1)	ATE(S = 0)		OM	IPW1	IPW2	DR1	DR2	DR3	MT
5,000	−1.0	1.0	−2.0	4.152	46.955	41.394	7.793	7.291	6.827	21.783
5,000	−2.0	−2.0	0.0	4.159	51.201	41.189	7.548	7.280	6.834	21.843
5,000	−2.0	−2.5	0.5	4.136	55.658	41.335	7.619	7.291	6.832	21.812
5,000	−2.0	−3.0	1.0	4.153	53.927	41.428	7.666	7.337	6.859	21.817
5,000	−2.0	2.0	−4.0	4.148	46.360	41.634	7.627	7.308	6.852	21.842

Abbreviations: ATE, average treatment effect; dATE, difference between the stratum-specific average treatment effects; DR, doubly robust; IPW, inverse probability weighting; MT, matching; OM, outcome modeling.

[a] Results for continuous outcome simulations with Pr[$S = 1$] = 0.5 and independent baseline covariates.

[b] ATE($S = s$), mean difference in stratum s.

[c] ATE($S = 1$) − ATE($S = 0$).

[d] Notation: OM, OM estimator in equation 4; IPW1, IPW estimator in equation 5; IPW2, IPW estimator in equation 6; DR1, DR estimator in equation 7; DR2, DR estimator in equation 8; DR3, DR estimator in equation 9; MT, MT estimator in equation 10.

[e] For each set of results by sample size (n), the fourth row has the same simulation parameters as the first row because setting ATE($S = 1$) = 0 leads to ATE($S = 0$) = 0, both when ATE($S = 0$) = ATE($S = 1$) and when ATE($S = 0$) = −ATE($S = 1$). Results may differ between the first and fourth rows due to simulation error.

followed by standardization, IPW, doubly robust weighted regression, and matching methods to adjust for confounding from the following baseline covariates: age, severity of angina, history of previous myocardial infarction, percent obstruction of the proximal left anterior descending artery, left ventricular wall motion score, number of diseased vessels, and ejection fraction. We chose these variables on the basis of a previous study that analyzed the same data (53), and we did not perform any model specification search for the outcome model or the propensity score (37). In the observational study, we also used an unadjusted analysis for comparison with the adjusted analyses to informally evaluate the impact of confounding.

Of the 2,093 patients in the CASS data set, 1,686 had complete data on all baseline covariates (731 randomized, 368 to surgery and 363 to medical therapy; 955 nonrandomized, 430 receiving surgery and 525 receiving medical therapy). For simplicity, we restricted our analyses to patients with complete data. In Web Appendix 6, Web Table 21, we give the baseline characteristics of patients contributing data to our analyses.

Model specification

In analyses of the observational component of CASS, we fitted logistic regression models for the probability of treatment and the probability of the outcome. We fitted the treatment model with the main effects of the baseline covariates, using restricted cubic splines for continuous covariates, and all possible 2-way interactions between the baseline covariates and the subgroup variable of interest (either ejection fraction or history of previous myocardial infarction). We fitted the outcome model with the main effects of these baseline covariates and treatment, the interaction between treatment and the subgroup variable of interest (either ejection fraction or history of previous myocardial infarction), and all possible 2-way interactions between the baseline covariates and the subgroup variable of interest.

Results

Estimates of the 10-year mortality risk and treatment effects at 10 years are shown in Tables 3 and 4 on the risk difference scale. We report results on the odds ratio scale in Web Appendix 6, in Web Tables 22 and 23, and the potential outcome means in Web Tables 24 and 25. We used bootstrap resampling (10,000 samples) to obtain percentile 95% confidence intervals for each estimator; normal-distribution–based bootstrap intervals were nearly identical and are not shown. Results in the observational data were similar across different estimators. Because different estimators rely on different working models—the IPW estimator relies on modeling the covariate-treatment association and the OM-based estimator relies on modeling the treatment/covariate-outcome association—agreement between them suggests that the results are not driven by modeling choices (54). The magnitudes of effect heterogeneity across subgroups defined by previous myocardial infarction were similar in the observational and randomized components of CASS (on both the odds ratio scale and the risk difference scale). The magnitude of effect heterogeneity across subgroups defined by ejection faction in the observational component was substantially smaller compared with the randomized component of CASS (on both scales, as indicated by the larger relative odds ratio and difference of risk differences in the trial vs. the observational analyses). The difference between the trial and observational analyses comparing patients within ejection fraction subgroups may be explained by lack of conditional exchangeability between the treated

Table 3. Subgroup Analysis of Previous Myocardial Infarction (S = 1 if History of Myocardial Infarction; S = 0 Otherwise) in the Coronary Artery Surgery Study, August 1975–December 1996

Estimator[a]	Effect in $S = 1$		Effect in $S = 0$		Comparison of Effects	
	RD[b]	95% CI[c]	RD[b]	95% CI[c]	DRD	95% CI[c]
Trial (unadjusted)	−4.43	−12.13, 3.37	1.36	−6.22, 8.89	−5.78	−16.50, 4.92
Obs (unadjusted)	−3.52	−10.39, 3.73	4.66	−2.11, 11.58	−8.18	−17.97, 1.52
OM	−4.25	−11.19, 3.07	4.84	−2.07, 11.99	−9.10	−19.09, 0.86
IPW1	−4.89	−11.95, 2.55	4.59	−2.09, 11.50	−9.49	−19.59, 0.39
IPW2	−4.82	−11.79, 2.61	4.40	−2.31, 11.29	−9.21	−19.07, 0.64
DR1	−4.89	−11.82, 2.58	4.48	−2.40, 11.56	−9.36	−19.30, 0.64
DR2	−4.88	−11.80, 2.58	4.48	−2.38, 11.56	−9.37	−19.30, 0.62
DR3	−4.84	−11.70, 2.56	4.45	−2.32, 11.56	−9.29	−19.17, 0.72
MT	−6.74	−12.93, 4.70	2.96	−2.97, 13.21	−9.70	−21.36, 2.61

Abbreviations: CI, confidence interval; DR, doubly robust; DRD, difference between the RDs; IPW, inverse probability weighting; MT, matching; Obs, observational; OM, outcome modeling; RD, risk difference.

[a] Notation: OM, OM estimator in equation 4; IPW1, IPW estimator in equation 5; IPW2, IPW estimator in equation 6; DR1, DR estimator in equation 7; DR2, DR estimator in equation 8; DR3, DR estimator in equation 9; MT, MT estimator in equation 10.

[b] Risk difference × 100 (% scale).

[c] 95 percent percentile CIs from 10,000 bootstrap resamples.

and untreated groups in the observational study (e.g., due to unmeasured confounding) or lack of exchangeability between the randomized and nonrandomized groups (e.g., due to differences in the distribution of effect modifiers other than ejection fraction) (55, 56).

Table 4. Subgroup Analysis for the Ejection Fraction (S = 1 if Ejection Fraction ≥ 50%; S = 0 Otherwise) in the Coronary Artery Surgery Study, August 1975–December 1996

Estimator[a]	Effect in $S = 1$		Effect in $S = 0$		Comparison of Effects	
	RD[b]	95% CI[c]	RD[b]	95% CI[c]	DRD	95% CI[c]
Trial (unadjusted)	1.89	−4.06, 7.75	−18.32	−32.02, −4.42	20.22	4.93, 35.15
Obs (unadjusted)	1.63	−3.31, 6.72	−9.15	−22.84, 5.00	10.78	−4.12, 25.52
OM	1.28	−3.78, 6.46	−12.06	−26.46, 2.19	13.35	−1.81, 28.76
IPW1	1.36	−3.80, 6.55	−11.11	−25.93, 4.87	12.46	−4.33, 28.26
IPW2	1.34	−3.84, 6.52	−10.81	−25.76, 4.74	12.14	−4.18, 28.19
DR1	1.28	−3.87, 6.47	−11.09	−25.70, 4.12	12.37	−3.75, 28.02
DR2	1.28	−3.87, 6.47	−11.08	−25.71, 3.99	12.36	−3.59, 28.02
DR3	1.31	−3.85, 6.51	−11.16	−25.75, 3.40	12.47	−2.85, 28.08
MT	2.32	−4.40, 8.53	−6.70	−29.01, 7.99	9.02	−6.90, 32.08

Abbreviations: CI, confidence interval; DR, doubly robust; DRD, difference between the RDs; IPW, inverse probability weighting; MT, matching; Obs, observational; OM, outcome modeling; RD, risk difference.

[a] Notation: OM, OM estimator in equation 4; IPW1, IPW estimator in equation 5; IPW2, IPW estimator in equation 6; DR1, DR estimator in equation 7; DR2, DR estimator in equation 8; DR3, DR estimator in equation 9; MT, MT estimator in equation 10.

[b] Risk difference × 100 (% scale).

[c] 95 percent percentile CIs from 10,000 bootstrap resamples.

DISCUSSION

Assessing heterogeneity of treatment effects is a key challenge in comparative effectiveness research (1, 2), but only a handful of studies have evaluated the performance of different methods for estimating subgroup-specific average treatment effects and examining whether these effects vary (9–16). Some investigators reported results from case studies using empirical data (9, 13, 14) and thus provided limited information about the performance of estimators; others only compared a limited number of competing methods in simulation studies (10–12, 15, 16).

We compared the performance of OM-based, IPW, doubly robust, and matching estimators of subgroup-specific potential outcome means, conditional average treatment effects, and differences of conditional average treatment effects for continuous and binary outcomes. We found that the bias and standard deviation of the sampling distribution of estimators varies substantially even under the best-case scenario of no confounding by unmeasured covariates and no model misspecification. When the working models for the expectation of the outcome and the probability of treatment were correctly specified, all estimators were nearly unbiased, even with fairly small sample sizes; there were, however, important difference in $\sqrt{n}$-scaled bias, which was much higher for IPW and matching estimators than for OM-based and doubly robust estimators. Furthermore, the OM-based estimator had the lowest $\sqrt{n}$-scaled standard deviation; the doubly robust estimators had a standard deviation larger than the OM-based estimator but much lower than that of the IPW or matching estimator.

In practice, we suggest using multiple estimators in order to assess whether model specification choices influence results (54). Combining outcome mean and probability-of-treatment models in doubly robust estimators is particularly attractive because it offers robustness against model specification and, as we discuss below, also has advantages when data-adaptive methods are needed. Normalizing the weights by their sum, both for IPW and for doubly robust estimators, usually leads to better performance overall (i.e., often the decreases in variance are more substantial than the increases in bias), especially when the sample size is large and weights are highly variable (28, 29). The performance of a simple matching estimator in our simulations was less than satisfactory; further research is needed to evaluate more refined estimation strategies that incorporate matching (e.g., see Colson et al. (57)) and to assess performance for different causal quantities (e.g., the conditional average treatment effect on the treated).

Our simulations used parametric regression models, fitted with standard maximum (quasi-) likelihood methods, because this is by far the most commonly used approach in applications and because we wanted to compare estimators under a best-case scenario. We also note that these fairly simple methods can form the basis for more sophisticated approaches for assessing heterogeneity of treatment effects—for example, for subgroup identification using tree-based methods with observational data (32). Nevertheless, further research is needed to examine the performance of the methods described in this paper when parametric models are misspecified (58).

To mitigate the risk of model misspecification, data-adaptive methods (e.g., machine learning methods) may be used to estimate the conditional expectation of the outcome or the probability of treatment (59–62). When using such methods, which typically converge to the true model at a rate slower than parametric, doubly robust estimators such as DR1 and DR2 can still converge at a $\sqrt{n}$ rate (31, 32).

ACKNOWLEDGMENTS

Author affiliations: Center for Evidence Synthesis in Health, School of Public Health, Brown University, Providence, Rhode Island, United States (Sarah E. Robertson, Andrew Leith, Christopher H. Schmid, Issa J. Dahabreh); Department of Health Services, Policy and Practice, School of Public Health, Brown University, Providence, Rhode Island, United States (Sarah E. Robertson, Issa J. Dahabreh); Department of Biostatistics, School of Public Health, Brown University, Providence, Rhode Island, United States (Christopher H. Schmid); Center for Statistical Sciences, School of Public Health, Brown University, Providence, Rhode Island, United States (Christopher H. Schmid); Department of Epidemiology, School of Public Health, Brown University, Providence, Rhode Island, United States (Issa J. Dahabreh); and Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States (Issa J. Dahabreh).

This work was funded by Patient-Centered Outcomes Research Institute (PCORI) awards ME-1306-03758 and ME-1502-27794 and Agency for Healthcare Research and Quality National Research Service Award T32AGHS00001.

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of PCORI, its Board of Governors, the PCORI Methodology Committee, or the Agency for Healthcare Research and Quality. The data analyses carried out in this paper used Coronary Artery Surgery Study (CASS) research materials obtained from the National Heart, Lung, and Blood Institute's Biologic Specimen and Data Repository Information Coordinating Center. This paper does not necessarily reflect the opinions or views of the CASS investigators or the National Heart, Lung, and Blood Institute.

Conflict of interest: none declared.

REFERENCES

1. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centered evidence. *Int J Epidemiol*. 2016;45(6):2184–2193.
2. Dahabreh IJ, Trikalinos TA, Kent DM, et al. Heterogeneity of treatment effects. In: Gatsonis C, Morton SC, eds. *Methods in Comparative Effectiveness Research*. Boca Raton, FL: Chapman & Hall/CRC Press; 2017:227–271.
3. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9):1393–1512.

4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
5. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48(12):479–495.
6. Scharfstein D, Rotnizky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models, with comments and rejoinder. *J Am Stat Assoc*. 1999; 94(448):1096–1146.
7. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4): 962–973.
8. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
9. Liem YS, Wong JB, Hunink MM, et al. Propensity scores in the presence of effect modification: a case study using the comparison of mortality on hemodialysis versus peritoneal dialysis. *Emerg Themes Epidemiol*. 2010;7(1):Article 1.
10. Rassen JA, Glynn RJ, Rothman KJ, et al. Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiol Drug Saf.* 2012;21(7):697–709.
11. Radice R, Ramsahai R, Grieve R, et al. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *Int J Biostat*. 2012;8(1):Article 25.
12. Kreif N, Grieve R, Radice R, et al. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Med Decis Making*. 2012;32(6):750–763.
13. Xie Y, Brand JE, Jann B. Estimating heterogeneous treatment effects with observational data. *Sociol Methodol*. 2012;42(1): 314–347.
14. Green KM, Stuart EA. Examining moderation analyses in propensity score methods: application to depression and substance use. *J Consult Clin Psychol*. 2014;82(5):773–783.
15. Eeren HV, Spreeuwenberg MD, Bartak A, et al. Estimating subgroup effects using the propensity score method: a practical application in outcomes research. *Med Care*. 2015; 53(4):366–373.
16. Wang SV, Jin Y, Fireman B, et al. Relative performance of propensity score matching strategies for subgroup analyses. *Am J Epidemiol*. 2018;187(8):1799–1807.
17. Varadhan R, Segal JB, Boyd CM, et al. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol*. 2013; 66(8):818–825.
18. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A*. 1984;656–666.
19. Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *J Am Stat Assoc*. 1994;89(427):737–749.
20. Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci*. 1990;5(4):465–472.
21. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688–701.
22. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
23. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41(2):514–520.
24. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20(6):880–883.
25. Hernán MA, Robins JM. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC Press; 2021.
26. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York, NY: Springer Publishing Company; 2000:1–94.
27. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
28. Hájek J. J. Hájek: [comment]. In: Godambe VP, Sprott DA, eds. *Foundations of Statistical Inference*. New York, NY: Holt, Rinehart, & Winston; 1971:236. (Comment on "An essay on the logical foundations of survey sampling, part one" by D. Basu (pp. 203–233)).
29. Robins J, Sued M, Lei-Gomez Q, et al. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat Sci*. 2007;22(4): 544–559.
30. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*. 1998;66(2):315–331.
31. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21(1):C1–C68.
32. Yang J, Dahabreh IJ, Steingrimsson JA. Causal interaction trees: tree-based subgroup identification for observational data [preprint]. *arXiv*. 2020. (doi: arXiv:2003.03042).
33. Tsiatis A. *Semiparametric Theory and Missing Data*. New York, NY: Springer Publishing Company; 2007.
34. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *J Econom*. 2007;141(2): 1281–1301.
35. Gourieroux C, Monfort A, Trognon A. Pseudo maximum likelihood methods: theory. *Econometrica*. 1984;681–700.
36. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J*. 2012;33(15): 1893–1901.
37. Ellis AG, Trikalinos TA, Wessler BS, et al. Propensity score–based methods in comparative effectiveness research on coronary artery disease. *Am J Epidemiol*. 2018;187(5): 1064–1078.
38. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262–270.
39. Leuven E, Sianesi B. Help for psmatch2. http://repec.org/bocode/p/psmatch2.html. Published 2003. Accessed August 17, 2020.
40. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006; 74(1):235–267.
41. Abadie A, Imbens GW. Matching on the estimated propensity score. *Econometrica*. 2016;84(2):781–807.
42. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29–38.
43. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19): 2937–2960.

44. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Stat Med*. 2014;33(5):721–737.
45. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. (Monographs on Statistics and Applied Probability, no. 57). Boca Raton, FL: Chapman & Hall/CRC Press; 1993.
46. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica*. 2008;76(6):1537–1557.
47. Otsu T, Rai Y. Bootstrap inference of matching estimators for average treatment effects. *J Am Stat Assoc*. 2017;112(520): 1720–1732.
48. Robins JM, Morgenstern H. The foundations of confounding in epidemiology. *Comput Math Appl*. 1987;14(9–12): 869–916.
49. Olschewski M, Scheurlen H. Comprehensive Cohort Study: an alternative to randomized consent design in a breast preservation trial. *Methods Inf Med*. 1985;24(3):131–134.
50. William J, Russell R, Nicholas T, et al. Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery. *Circulation*. 1983;68(5):939–950.
51. CASS Principal Investigators. Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery. Comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria. *J Am Coll Cardiol*. 1984;3(1): 114–128.
52. Chaitman BR, Ryan TJ, Kronmal RA, et al. Coronary Artery Surgery Study (CASS): comparability of 10 year survival in randomized and randomizable patients. *J Am Coll Cardiol*. 1990;16(5):1071–1078.
53. Olschewski M, Schumacher M, Davis KB. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Control Clin Trials*. 1992;13(3):226–239.
54. Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: some questions and an answer". *Stat Sin*. 2001;11(4):920–936.
55. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. *Eur J Epidemiol*. 2019;34(8):719–722.
56. Dahabreh IJ, Robertson SE, Steingrimsson JA, et al. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020;39(14):1999–2014.
57. Colson KE, Rudolph KE, Zimmerman SC, et al. Optimizing matching and analysis combinations for estimating causal effects. *Sci Rep*. 2016;6:23222.
58. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4): 523–539.
59. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3): 337–346.
60. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol*. 2010; 63(8):826–833.
61. Watkins S, Jonsson-Funk M, Brookhart MA, et al. An empirical comparison of tree-based methods for propensity score estimation. *Health Serv Res*. 2013;48(5):1798–1817.
62. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546–555.