Published in final edited form as:
Stat Med. 2013 September 10; 32(20): 3552–3568. doi:10.1002/sim.5802.

The Use of Propensity Scores and Observational Data to Estimate Randomized Controlled Trial Generalizability Bias

Taylor R. Pressler[a] and **Eloise E. Kaizar**[b,*]
[a]*Quantum Health, Columbus, Ohio 43235, U.S.A.*
[b] Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A.

Abstract

While randomized controlled trials (RCT) are considered the "gold standard" for clinical studies, the use of exclusion criteria may impact the external validity of the results. It is unknown whether estimators of effect size are biased by excluding a portion of the target population from enrollment. We propose to use observational data to estimate the bias due to enrollment restrictions, which we term generalizability bias. In this paper we introduce a class of estimators for the generalizability bias and use simulation to study its properties in the presence of non-constant treatment effects. We find the surprising result that our estimators can be unbiased for the true generalizability bias even when all potentially confounding variables are not measured. In addition, our proposed doubly robust estimator performs well even for mis-specified models.

Keywords

Observational studies; Randomized controlled trials; Sample selection error; Propensity score; Causal effect

1. Introduction

In many areas of clinical research, interest lies in quantifying the treatment effect of an intervention. In general, randomized controlled trials (RCTs) are considered the "gold standard" for such evaluations. This status is due to the strong internal validity derived from treatment randomization. However, these studies may have limited external validity since the study participants may represent only a small segment of the population of interest. Concerns about external validity arise from a mismatch between study design and widespread study use. Many trials are designed to demonstrate treatment efficacy. That is, many trials are designed as a 'proof of concept' to show that the intervention improves an outcome on average for some part of the population.

Toward this aim, RCTs are often designed to demonstrate a statistically significantly large effect size, while minimizing potential risks to patients and costs to trial sponsors. This means that those who are *a priori* thought to benefit most from the intervention are actively recruited through inclusion criteria, and those who are *a priori* thought to be at the greatest risk or to notably increase the cost of the RCT are actively excluded through exclusion criteria. In particular, those with severe forms of disease or multiple co-morbidities, women of childbearing age, and children are often excluded from RCTs [1].

[*] Correspondence to: Eloise Kaizar, Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A. ekaizar@stat.osu.edu.

Despite the clear efficacy designs of many RCTs, US FDA regulation and evidence-based practice promote generalizing trial results to inform decisions for members of the entire population. The validity of generalizing RCT results has been studied from the standpoint of the size of the excluded population. Studies across a range of fields have shown that potentially important characteristics vary significantly across inclusion groups [2] and that it is not unusual for *less than half* of an affected population to be eligible for RCT participation (see, for example, [3, 4, 5, 6]). In part due to these studies, some practitioners may distrust and underuse RCT evidence [7]. Conversely, some practitioners may place too much trust in RCT evidence, disregarding the potential importance of the inclusion and exclusion criteria for the trial participants and even omitting these criteria from their own reports [8].

A large excluded population is only one criterion for trial results to be biased estimates of effectiveness in the broader population. The treatment effect size must also differ between those who are and are not eligible for trial participation. If the effect of treatment is indeed constant over the population of interest, RCT-population unbiased estimators would also be unbiased for any population because the designed exclusion would have no impact on the effect measured in the study. However, as the recent interest in personalized medicine evidences, many treatments will have non-constant effects. Variability of treatment effect size across exclusion criteria has, to our knowledge, not been extensively studied in the literature. Direct examination of treatment effect heterogeneity can appropriately reassure practitioners that RCT results are widely applicable, or raise concerns about using RCT evidence to inform decisions about excluded populations. One way to investigate treatment effect heterogeneity is to consider the bias incurred by using an RCT-derived estimator to make inference about a broader population of interest, or an RCT's generalizability bias.

By studying generalizability bias among similar studies, we can either justify or bring a note of caution to the current use of RCT-based estimates for broad implementation of evidence-based medicine in that field. By examining generalizability bias across a wide range of studies, we can provide specific evidence for broad improvements in study design that would minimize generalizability bias. Current research into the adequacy of RCT design (e.g., [1, 3, 4, 5, 6]) demonstrates the interest in the generalizability of RCT results, but is limited to raising concerns about the potential for generalizability bias. Our proposed methods allow the empirical estimation of generalizability bias to assess the external validity of completed trials that does not require access to raw trial data. In some cases, our methods could be used to predict the generalizability bias during the planning stages of a trial and can thus improve its design.

Estimates of generalizability bias also have a second important use. Several authors have proposed methods for incorporating bias into estimates of population average treatment effects (PATEs) [9, 10, 11]. Our methods provide empirical evidence for the magnitude of these biases that can be directly utilized in these types of estimators. Large-scale decision-makers, such as those who set insurance formularies, still make decisions for broad populations. Thus, PATEs for the whole population for which decisions are made is of interest, even as some research turns to more patient-centered analyses. Furthermore, we believe that our work in developing generalizability estimates for broad populations can in the future be incorporated in the construction of unbiased estimates for average treatment effects in specifically defined subpopulations, such as those used in patient-centered outcomes research.

The literature already contains several methods for considering the generalizability of RCTs. For example, Greenhouse, et al describe a formal approach to judging the generalizability of RCTs by comparing their participants to population characteristics estimated from observational data [2]. Marcus evaluates the generalizability of RCT results by directly comparing RCT participants to those who were not eligible to be randomized but were followed in a registry [12]. Both of these approaches are valuable for certain types of data availability. The former is useful when treatment choice is difficult to measure or assess in observational studies. The latter relies on registries of unenrolled patients that are unfortunately rarely funded. Several authors have also used expert knowledge to accommodate generalizability bias in their estimation via Bayesian models with prior distributions on generalizability bias parameters. See, e.g. [11, 13].

In contrast to relying on expert opinion, we propose a class of estimators of generalizability bias based on empirical evidence. Since randomized evidence is not typically available for estimating the treatment effect among the excluded population and by definition can not directly provide information about the generalizability bias, our approach relies entirely on observational data, such as can be derived from electronic medical records. Our estimators combine estimates of the average effect size in both the included and excluded portions of the population of interest. Administrative data are becoming more accessible and useful for examining effectiveness in natural settings and among wide populations. These data are especially applicable for assessing interventions that have been available for some time, as for phase 4 pharmaceutical and medical device studies.

The use of observational data to estimate average effect size has been the subject of much recent statistical interest. Of greatest concern is the issue of weak internal validity due to imbalanced treatment selection. Our proposed estimators account for treatment selection via well known propensity score based methods. (See, e.g., [14, 15] for an overview of propensity score methods.) However, because we take the difference between effect estimates, we can relax the usual requirements of propensity-based estimation. In particular, in order to create an unbiased estimate of the generalizability bias, we need only assume that the bias of the average effect size estimators for the included population is the same as the bias of the average effect size estimator for the excluded population. We have found plausible data situations in which this condition holds, leading to unbiased estimation of the generalizability bias of RCTs, and explain these situations in detail in Section 3.

We demonstrate the use of our proposed estimators of generalizability bias using a real example. Williams, et al. conducted an RCT in part to determine the effect of using a vacuum device vs. the more traditional forceps in an assisted birth on maternal outcomes [16]. However, as is common in RCTs, this study excluded young women. We use this exclusion to demonstrate our proposed methods to estimate the bias in the estimated effects on maternal length of stay and severe laceration.

Our paper is organized as follows. In Section 2, we review the framework for our problem, including the causal framework, as well as common causal estimators. In Section 3, we place generalizability bias within the causal framework, present our proposed estimators and note the ideal data situations for unbiased estimation. We describe the simulation studies that demonstrate the small-sample performance of our estimators in Section 4. We present a real data analysis in Section 5, and finally discuss the implications of our results in Section 6.

2. Common Causal Estimands and Estimators

2.1. Causal Structure

The goal of a study of a treatment is to estimate its effect. For example, consider a study to compare maternal outcomes following the use of a vacuum device vs. the more traditional forceps in an assisted birth. The ideal study would estimate the difference in outcomes for a specific mother facing this birth complication, if the delivery were to be assisted by the control or active interventions, which we take to be forceps or vacuum, respectively. We denote these potential outcomes for subject *i* by $Y_i(0)$ and $Y_i(1)$, respectively. The treatment effect for a single subject is thus $Y_i(1) – Y_i(0)$. Because we can not observe both control and active interventions for any one birth, we instead estimate an average treatment effect.

We take the treatment effect of interest to be the population average treatment effect (PATE), denoted Δ. PATE is the average treatment effect over all N individuals in a finite population of interest:

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1) - Y_i(0)\right] \quad (1)$$

or the expected treatment effect in an infinite population of interest:

$$\Delta = E\left[Y_i(1) - Y_i(0)\right] = E\left[Y_i(1)\right] - E\left[Y_i(0)\right]$$

Without loss of generalizability, we consider the finite population PATE.

2.2. Common Estimators

Our estimates of generalizability bias that we present in detail in Section 3 are built upon estimates of average treatment effect for two specific subpopulations based on the analysis of a large observational dataset, such as an administrative database. Regardless of how we define the population of interest (i.e., the whole population or a subpopulation), estimating average treatment effects using observational data can be a challenge. While many methods have been proposed, we consider four popular methods of estimating Δ appropriate for observational data sets: a simple estimator with no covariate adjustments, a regression estimator, an inverse propensity weighted (IPW) estimator, and a doubly-robust (DR) estimator. Each estimate is calculated from a simple random sample of size *n* from the whole population, denoted by the set of indices ω. Each estimate is a consistent estimate of Δ under certain data configurations, as detailed below and outlined in Table 1.

For the remainder of this paper, we assume that a large observational dataset is available for analysis, and that this dataset approximates a simple random sample from the population of interest. Our estimate of the generalizability bias utilizes only the design information of an RCT, since by design the results of an RCT can not provide information on the generalizability bias.

2.2.1. Simple Estimator—A simple estimator of PATE is a simple difference of averages, or in the language of survey sampling, a difference of estimated domain averages:

$$\widehat{\Delta}_{sim} = \frac{1}{n_1}\sum_{i\in\omega} Y_i T_i - \frac{1}{n_0}\sum_{i\in\omega} Y_i\left[1 - T_i\right] \quad (2)$$

where n_1 and n_0 are the number of sampled individuals who receive active and control intervention, respectively, and T_i is an indicator of active treatment receipt. Elementary survey sampling theory tells us that non-overlapping domain estimates are uncorrelated, and so, for large samples, the variance of the simple estimator of PATE can be estimated by:

$$\widehat{V}\left(\widehat{\Delta}_{sim}\right) = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}, \quad (3)$$

where s_i is the sample standard deviation of the outcome among those that receive treatment *i*, for *i* = 0, 1. If the population size *N* is not much larger than the sample size *n*, a finite population correction can be added to improve the variance estimate. Again, based on survey sampling theory, we know that the simple estimator is an unbiased estimator of PATE if our sample and treatment selection together represent a two-stage simple random sample (SRS). A sufficient condition for unbiasedness is that each individual's choice of intervention is independently and identically distributed, (i.e., there is no confounding).

2.2.2. Regression Estimator—We can also construct an estimate by modeling the potential outcomes conditional on variables that confound the effect of treatment choice and treatment itself, **X**. Although more complex models are feasible (e.g., [17]), we follow usual practice by using a linear model:

$$Y\left(\mathbf{X}_i, T_i\right) \overset{ind}{\sim} N\left(\mu\left(\mathbf{X}_i, T_i, \beta\right), \sigma^2\right).$$

Based on this model, a regression-based estimator of PATE is:

$$\widehat{\Delta}_{reg} = \frac{1}{n_1}\sum_{i\in\omega}\left[\widehat{\mu}\left(\mathbf{X}_i, T=1, \widehat{\beta}\right) - \widehat{\mu}\left(\mathbf{X}_i, T=0, \widehat{\beta}\right)\right], \quad (4)$$

where $\widehat{\mu}\left(\mathbf{X}_i, T=t, \widehat{\beta}\right)$ is the predicted mean for individual *i* from a linear regression model at the observed covariate values $\mathbf{X}_i$ had that individual received treatment *t*. Relying on the usual Normal iid error assumptions, we estimate the variance to be:

$$\widehat{V}\left(\widehat{\Delta}_{reg} = \mathbf{1}(\boldsymbol{Z}_{\mathbf{1}} - \boldsymbol{Z}_{\mathbf{0}})'\widehat{\mathbf{V}}\left(\widehat{\beta}\right)(\boldsymbol{Z}_{\mathbf{1}} - \boldsymbol{Z}_{\mathbf{0}})\,\mathbf{1}, \quad (5)$$

where $\widehat{V}\left(\widehat{\beta}\right)$ is the usual estimator of the variance of the estimated coefficients, and $\mathbf{Z}_t$ is the regression model matrix if the entire subsample had received treatment t. The regression estimator is consistent if the regression model is correct and the coefficients β estimable.

2.2.3. Inverse Propensity Weighted Estimator—Another approach to overcoming the influence of treatment choice is to adjust the estimates using the estimated probabilities of receiving the active treatment, called propensity scores [14, 15]. Following the ideas of survey statistics, inverse propensity weighted (IPW) estimation reweights the entire population to consistently estimate the average counterfactual for treatment and the average counterfactual for control, and subtracts the two:

$$\widehat{\Delta}_{ipw} = \widehat{\mu}\left(T=1\right) - \widehat{\mu}\left(T=0\right) = \left[\sum_{i\in\omega}\frac{T_i}{\widehat{e}_i}\right]^{-1}\sum_{i\in\omega}\left[\frac{Y_iT_i}{\widehat{e}_i}\right] - \left[\sum_{i\in\omega}\frac{1-T_i}{1-\widehat{e}_i}\right]^{-1}\sum_{i\in\omega}\left[\frac{Y_i\left[1-T_i\right]}{1-\widehat{e}_i}\right], \quad (6)$$

where the weights are based on the estimated propensity scores, $\hat{e}(\mathbf{X})$. We estimate these via a logistic regression of observed treatment on a collection of relevant covariates $\mathbf{X}_i$:

$$\begin{array}{rl} T_i|\mathbf{X}_i, \alpha \overset{\mathbf{iid}}{\sim} & \text{Bernoulli}\left(e\left(\mathbf{X}_i\right)\right) \\ \text{logit}\left[e\left(\mathbf{X}_i\right)\right] = & \mathbf{X}_i\alpha \end{array} \quad (7)$$

Generally, if the propensity model is correct and there is sufficient overlap in the propensity scores of the treatment and control groups, the IPW estimator is consistent. Rosenbaum and Rubin present detailed conditions for consistency [18]. Lunceford and Davidian present an estimator of the variance of the IPW estimator in their Equation 19 [19]. It is well known that the variance of IPW estimators can grow quite large whenever the inverse propensity scores grow large, and thus alternative estimators, or those that introduce some bias by truncating the propensity scores, may be more desirable.

2.2.4. Doubly Robust Estimator—

While the regression and IPW estimators are based on only the response and propensity models, respectively, they may be sensitive to model mis-specifications. Doubly robust estimators, $\widehat{\Delta}_{dr}$, are robust to mis-specification of either the propensity or response models. While there are many forms of doubly robust estimators, we use one presented by Lunceford and Davidian [19]:

$$\widehat{\Delta}_{dt}=\frac{1}{n}\sum_{i\in\omega}\frac{Y_iT_i-[T_i-\widehat{e}_i]\,m_1\left(\mathbf{X}_i,\widehat{\beta}_1\right)}{\widehat{e}_i}-\frac{1}{n}\sum_{i\in\omega_j}\frac{Y_i\,[1-T_i]+[T_i-\widehat{e}_i]\,m_0\left(\mathbf{X}_i,\widehat{\beta}_0\right)}{1-\widehat{e}_i},\quad(8)$$

where $m_t\left(\mathbf{X}_i,\widehat{\beta}_t\right)=\widehat{E}\,[Y|T=t,\mathbf{X}_i]$ is the predicted response for individual *i* from a regression of the response **Y** on the covariates **X** and $\widehat{\beta}_t$ are estimated separately for $t = 0,1$, using only the individuals in the control group or treatment group, respectively. While on its face, this estimator is not intuitive, Kang and Schafer show that it can be rewritten as a regression-based estimator with a propensity-based correction for response model mis-specification [20]. Lunceford and Davidian present an estimator of the variance of the IPW estimator in their Equation 21 [19].

3. Generalizability Bias

3.1. Estimand

Due to explicit exclusion criteria for the RCT, the population of interest is split into two subpopulations. The first includes all N_I individuals who would be eligible for inclusion in the RCT; the second includes the remaining $N - N_I = N_E$ individuals. The average treatment effect for each of these is called the subpopulation average treatment effect (SPATE), and denoted $\Delta(j)$, $j = I, E$:

$$\Delta\,(I)=\frac{1}{N_I}\sum_{i=1}^{N}\left(Y_i\,(1)-Y_i\,(0)\right)I_i\quad(9)$$

$$\Delta\,(E)=\frac{1}{N_E}\sum_{i=1}^{N}\left(Y_i\,(1)-Y_i\,(0)\right)(1-I_i)\quad(10)$$

where I_i, equals 1 if individual *i* is eligible for inclusion in the trial and equals 0 otherwise.

If the average treatment effect in the whole population equals that in the RCT-eligible subpopulation, i.e., $\Delta = \Delta(I)$, then unbiased estimates of $\Delta(I)$ derived from the RCT data are also unbiased estimates of the estimand of interest, Δ, and there is no concern about generalizability. However, in many situations treatment effect heterogeneity in the population is such that $\Delta \neq \Delta(I)$. In this case, we must be concerned with the *generalizability bias* that results from estimating Δ with an unbiased estimator of $\Delta(I)$, denoted γ:

$$\begin{aligned}\gamma= & \Delta(I)-\Delta \\ = & \Delta(I)-\frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1)-Y_i(0)\right) \\ = & \Delta(I)-\frac{N_I}{NN_I}\sum_{i=1}^{N}\left(Y_i(1)-Y_i(0)\right)I_i-\frac{N_E}{NN_E}\sum_{i=1}^{N}\left(Y_i(1)-Y_i(0)\right)(1-I_i) \\ = & \Delta(I)-\frac{N_I}{N}\Delta(I)-\frac{N_E}{N}\Delta(E) \\ = & \pi_E\left(\Delta(I)-\Delta(E)\right],\end{aligned} \quad (11)$$

where $\pi_E=\frac{N_E}{N}$, the proportion of the population ineligible for RCT participation. In the following section, we propose a class of estimators for γ based on an available relevant observational dataset.

3.2. Estimators

Because the generalizability bias can be represented as a weighted average of the two SPATEs (Equation 11), a natural estimator is the weighted average of two estimates of the SPATEs:

$$\hat{\gamma}=\hat{\pi}_E\left[\hat{\Delta}(I)-\hat{\Delta}(E)\right]$$

Our proposed estimators rely entirely on observational data to estimate each component of the generalizability bias estimator. By not simultaneously using randomized data, we avoid potential biases due to other differences between randomized and observational studies, as we discuss further in Section 6. In addition, our method does not require access to RCT data or results, and can even be used in study design.

We consider the observational data to approximate an SRS of size *n* from the population of interest. That is, we assume the probability of selection for each individual does not depend on the treatment received or any other characteristic. For example, suppose a hospital wishes to estimate the generalizability bias or PATE for its typical patients. The patients treated in that hospital in recent years is reasonably considered to be an approximate random sample from the population of patients treated there in the next several years. Thus, it is also reasonable to estimate the proportion of individuals that would be excluded from a trial with the observed proportion: $\hat{\pi}_E=\frac{n_E}{n}$, where n_E is the number of individuals in the sample that would be ineligible for or excluded from trial participation. What remains is to estimate the SPATE for each segment of the population, $\hat{\Delta}(j), j=I, E$. Our general procedure is to divide the observational sample into those who would be excluded from trial participation, denoted by the set ω_E, and those who would be eligible for trial participation, denoted by the set ω_I. Then, we estimate the SPATE separately for each subpopulation using the appropriate subsample and one of the estimators presented in Section 2.2.

Regardless of the SPATE estimator that is employed, we take advantage of the approximate independence of the estimators in the two strata [21] to find an expression for the variance of the generalizability bias:

$$\begin{aligned}V\left[\gamma\right]= & \ V\left[\widehat{\pi}_{E}\left[\widehat{\Delta}\left(I\right)-\widehat{\Delta}\left(E\right)\right]\right]\\ = & \ V_{\widehat{\pi}_{E}}\left[E\left[\widehat{\pi}_{E}\left[\widehat{\Delta}\left(I\right)-\widehat{\Delta}\left(E\right)\right]\left|\widehat{\pi}_{E}\right]\right]+E_{\widehat{\pi}_{E}}\left[V\left[\widehat{\pi}_{E}\left(\widehat{\Delta}\left(I\right)-\widehat{\Delta}\left(E\right)\right]\left|\widehat{\pi}_{E}\right]\right]\\ \approx & \ V_{\widehat{\pi}_{E}}\left[\widehat{\pi}_{E}\left[\Delta\left(I\right)-\Delta\left(E\right)\right]\right]+E_{\widehat{\pi}_{E}}\left[\widehat{\pi}_{E}^{2}\right]V\left[\widehat{\Delta}\left(I\right)-\widehat{\Delta}\left(E\right)\right]\\ \approx & \ \tfrac{1}{n}\pi_{E}\left(1-\pi_{E}\right)\left[\Delta\left(I\right)-\Delta\left(E\right)\right]^{2}+\left[\tfrac{\pi_{E}\left(1-\pi_{E}\right)}{n}+{\pi_{E}}^{2}\right]\left[V\left[\widehat{\Delta}\left(I\right)\right]+V\left[\widehat{\Delta}\left(E\right)\right]\right]\\ = & \ \tfrac{\left(1-\pi_{E}\right)}{n\pi_{E}}\gamma^{2}+\left[\tfrac{\pi_{E}\left(1-\pi_{E}\right)}{n}+{\pi_{E}}^{2}\right]\left[V\left[\widehat{\Delta}\left(I\right)\right]+V\left[\widehat{\Delta}\left(E\right)\right]\right]\end{aligned} \qquad (12)$$

In this derivation we also assume that the estimator is approximately unbiased, and the contribution of the proportion of excluded population ($\widehat{\pi}_{E}$) to the variability of the SPATE estimators is negligible. In practice, the impact of these assumptions is unlikely to be meaningful, as they relate to variability terms of higher order than the main drivers of the variability – the variance of the SPATE estimates. We use the plug-in estimator for the variance:

$$\widehat{V}\left[\gamma\right]=\frac{\left(1-\widehat{\pi}_{E}\right)}{n\widehat{\pi}_{E}}\widehat{\gamma}^{2}+\left[\frac{\widehat{\pi}_{E}\left(1-\widehat{\pi}_{E}\right)}{n}+\widehat{\pi}_{E}^{2}\right]\left[\widehat{V}\left(\widehat{\Delta}\left(I\right)\right)+\widehat{V}\left(\widehat{\Delta}\left(E\right)\right)\right], \qquad (13)$$

and a Normal distributional approximation to calculate confidence intervals.

We have identified two sets of sufficient conditions where $\widehat{\gamma}$ is a consistent estimator of γ. First, $\widehat{\gamma}$ is consistent if each of its components, $\widehat{\Delta}\left(I\right)$ and $\widehat{\Delta}\left(E\right)$, is also consistent. This occurs when the appropriate conditions are met, as noted in Section 2.2 and summarized in Table 1. Generally speaking, the simple estimator is consistent if there is no confounding. If the response model is correctly specified, the regression and doubly-robust estimators are consistent. If the propensity model is correct, the IPW and doubly robust estimators are consistent. In any of these cases, the corresponding estimator of generalizability bias is also consistent for the true generalizability bias. Of course, these conditions can not be established directly from the observational data. In particular, the propensity-based estimators rely on the assumption of no unmeasured variables that would confound the SPATE estimates. Our estimator of generalizability bias benefits from an additional condition for consistency. If each estimate of SPATE is indeed biased, but this bias is identical in each subpopulation, then the overall estimate of the generalizability bias will remain consistent. We further examine this possibility focusing on the IPW estimator.

Extending previous results, Li, et al suggest that the bias of the IPW estimator due to uncontrolled confounding within each subpopulation stratum, *j* = *I*, *E*, is approximately [22]:

$$B_{\Delta(j)}=\frac{1}{N_j}\sum_{i\in\Omega_j}\left[\left\{E\left[Y_i\left(1\right)\left|T_i{=}1,\widehat{e}_i\right.\right]-E\left[Y_i\left(1\right)\left|T_i{=}0,\widehat{e}_i\right.\right]\right\}P\left[T_i{=}0|\widehat{e}_i\right]+\left\{E\left[Y_i\left(0\right)\left|T_i{=}1,\widehat{e}_i\right.\right]-E\left[Y_i\left(0\right)\left|T_i{=}0,\widehat{e}_i\right.\right]\right\}P\left[T_i{=}0|\widehat{e}_i\right]\right],$$

where $\hat{e}_i$ is the propensity score estimated from the measured variables and possibly incorrect propensity model, and T_i is an indicator of receipt of active treatment. Obviously, if the sufficient conditions for consistency hold (see above and Table 1) for both the RCT-eligible and -ineligible populations, there is no uncontrolled confounding and the terms of Equation 14 in curly braces all equal zero. However, our estimator of γ is proportional to the *difference* between the estimators of SPATE in the two strata. Thus, the bias of $\widehat{\gamma}$ will also be small if the bias of the two estimators of SPATE are not exactly zero, but approximately equal. That is, if the model mis-specification in the two strata results in estimators with large but approximately the same bias, $B_{\Delta(I)} \approx B_{\Delta(E)}$, then $\widehat{\gamma}$ will still be approximately unbiased.

That is, there are two sets of sufficient conditions for an IPW-based estimator of generalizability bias (i.e., the generalizability bias estimators based on the IPW and DR estimators) to be unbiased:

A. $\widehat{\Delta}(I)$ and $\widehat{\Delta}(E)$ consistent (see Table 1)

B. Bias of $\widehat{\Delta}(I) = \text{Bias} \quad \text{if} \quad \widehat{\Delta}(E)$, which occurs whenever the following three conditions are met:

  1. $I \perp\!\!\!\perp U|T, X$,

  2. $I \perp\!\!\!\perp U/X$ (or $I \perp\!\!\!\perp T|U, X$, or $T \perp\!\!\!\perp U|I, X$ for binary U).

  3. $E[Y|T, I, X, U] = g_1(T, I, X) + g_2(U, X)$,

where the notation $X \perp\!\!\!\perp Y|Z$ indicates the independence of X and Y conditional on Z. Conditions (B) are suggested by Li, et al [22] and Kaizar (2011) [10], who considers the bias of $\widehat{\gamma}$ in the case of the simple estimator. When the unmeasured variable U is binary, condition (B2) can be replaced with identical condition (B2*) $I \perp\!\!\!\perp T/U, X$ or (B2**) $T \perp\!\!\!\perp U|I, X$.

Heuristically, conditions (B1) and (B2) mean that the unmeasured variable U and choice of treatment T can only directly affect (or be affected by) the RCT exclusion critera I through dependence with covariates X. (For graphical models, conditions (B1) and (B2) imply that I is separated from T and U by X.) Further, condition (B3) means that the unmeasured variable U can not interact with the choice of treatment T or the exclusion criteria I to predict the observed outcome Y. For example, if the response model is linear, condition (B3) implies no UI or UT interaction terms, such as in the model:

$$\begin{aligned} Y_i|X_i, U_i, T_i, I_i = & \quad \beta_0+\beta_1 X_i+\beta_2 U_i+\beta_3 T_i+\beta_4 I_i+\beta_5 X_i U_i+\beta_6 X_i T_i+\beta_7 X_i I_i+\epsilon_i \\ \epsilon_i \overset{iid}{\sim} & \quad N\left(0, \sigma^2\right). \end{aligned}$$

Because the unmeasured variables are in practice unknown, we cannot empirically assess these conditions or the unbiasedness of our estimates. We demonstrate the doubly-unbiased advantage of our estimator in the second simulation study presented in Section 4, but use the assisted delivery example to clarify the practical assessment of these conditions. Recall that pregnant women at least 18 years old (adults) were eligible to participate in the trial, and the number of days the mother remained in the hospital is the response. Suppose that the number of previous deliveries, or parity, is an important unmeasured variable. If the parity influences the choice of treatment and the length of stay, estimates of PATE (both overall and among the included subpopulation) that do not control for parity will be biased. However, since it is reasonable that (B1) U=parity is unrelated to l=adult status conditional on X=age and T=treatment choice, and (B2*) the T=treatment choice is unrelated to I=adult status conditional on U=parity and X=age. Finally, while U=parity may interact with X=age to determine average Y=length of stay, it is unlikely to interact with I=adult status or T=treatment in predicting the Y=length of stay, satisfying condition (B3). Thus, our method should provide an approximately unbiased estimate of the generalizability bias in this study, even though parity may confound estimates of PATE.

Another example of when our method may work well is the exclusion of patients deemed at risk for suicide, or displaying ‘suicidality’ at the start of a study of a mental health treatment. Suppose the underlying severity of the illness is an important but unmeasurable variable, but each patient's illness evaluation (e.g., Hamilton rating scale score, HRSS) is available in an electronic health record. It is reasonable that U=severity influences X=HRSS, which in turn is used to determine I=suicidality and in practice influence choice of T=treatment. Thus, U=severity is independent of I=suicidality both (B1) conditional on X=HRSS and T=treatment, and (B2) conditional on X=HRSS alone. Finally, condition (B3) is satisfied if the outcome measure (e.g., 8-week post-treatment HRSS score) depends on a main effect of U=severity and the interaction between pre-score and treatment (e.g., the observed effect of treatment is greater for those who are measured to be more ill at the start of treatment).

However, one could posit cases when our estimators would be biased. Consider a randomized trial of the effect of inhaled corticosteroids on pulmonary function among patients with asthma. Many such trials exclude patients who smoke cigarettes [6]. It is also thought that cigarette smoke leads to corticosteroid resistance [23], which would result biased RCT estimation for any trial that excludes smokers. Suppose the observational data did not contain any covariates that measured U=airway damage, which could be due to smoking or a number of other also unmeasured causes, and which is also an important variable in the model for pulmonary function. Thus, the important unmeasured variable is not independent of the stratum indicator conditional on the treatment and measured covariates, violating condition (B1). If airway damage is also related to treatment choice and Y=pulmonary function (and thus potentially confounds the effect estimation), our methods may result in inconsistent estimation of the generalizability bias. The bias in estimating the effect of steroids due to unmeasured airway damage among smokers is not equivalent to the corresponding bias among non-smokers. This difference in biases is not detectable in the observational data, and will result in biased estimates of generalizability bias.

4. Simulation Study

We conducted two separate simulation studies to demonstrate two separate features of our estimators. First, we show that our estimators of generalizability bias are consistent whenever the SPATE is unbiasedly estimated within each substratum (sufficient condition A). Second, we show that the sufficient conditions B discussed in Section 3.2 lead to consistent estimation of generalizability even when SPATE is not consistently estimated within strata.

4.1. Design

For both simulation studies, we emulated a very simplified version of data that is commonly available in practice. We consider samples of size 10,000, which is consistent with the size of a small administrative database or a subset of a large database defined by a rare disorder. In each case, we consider three covariates, a binary treatment, and a continuous response variable.

For each study, the following sampling and estimation procedures were repeated 1,000 times. Each iteration consists of the following:

- A sample of size 10,000 is generated according to the models described below.
- Point estimates and intervals are calculated for SPATE in both subpopulations using each estimator $\widehat{\Delta}(j), j=I, E$ described in Section 2.2. Normal approximations are used to calculate each confidence interval.
- Point estimates and intervals are calculated for the generalizability bias using each estimator $\widehat{\gamma}$ described in Section 3.2.

4.1.1. Unbiased $\widehat{\Delta}(j)$, Unbiased $\widehat{\gamma}$—We generate three continuous covariates via a multivariate normal random number generator each with mean zero and variance of one. We assumed a uniform correlation among the three covariates that ranges from zero to one. We assigned treatment according to a logistic regression model that included all three continuous covariates as described in Equation 7. The intercept in the logistic regression was set to –0.7 and the coefficients of each covariate were set to 0.5. These values of α were chosen so that approximately 30% of the population elected treatment. The response variable is simulated according to a regression model with constant variance of one and mean:

$$E\left[Y_i|\mathbf{X}_i, T_i\right] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 T_i + \beta_5 X_{i3} T_i. \quad (15)$$

The interaction between X_3 and treatment means that the treatment effect is not constant across individuals in the population. For simplicity and to keep the effect of each coefficient comparable, each coefficient is set equal to one: $\beta_j = 1$, $j = 0 \ldots 5$. According to this model, the true PATE is $\beta_4 + \beta_5 E[X_{i3}] = 1$.

Finally, we imposed a trial exclusion criterion based on the variable X_1, which does not directly interact with treatment in the response model. Thus, if the exclusion variable X_1 is independent of the interaction variable X_3, the SPATE for the subpopulations are equal to the overall PATE. As the correlation between these two variables increases, the magnitude of the generalizability bias also increases, since SPATE is $\beta_4 + \beta_5 E[X_{i3}|X_{i1} < c]$, where c is the exclusion cutoff value. Figure 1 displays the change in SPATE(I) and generalizability bias as the correlation among the X variables ranges from 0 to 0.9, for cutoffs that exclude 25%, 50%, and 75% of the population as determined via normal quartiles.

We use these same treatment and response models for our generalizability bias estimators. In this case, we expect each of our estimators to be unbiased. However, we are also concerned with the effect of mis-specified models, namely the impact of unmeasured confounding. To explore this, we omit the variable X_3 from the model to estimate the propensity scores, the model for the outcome, or both of the models simultaneously. Our response model obviously does not meet sufficient condition (3), and so we expect our estimators to be biased for the true generalizability bias, γ.

4.1.2. Biased $\widehat{\Delta}(j)$, Unbiased $\widehat{\gamma}$—To clearly simulate data structured according to the three sufficient conditions for unbiased estimation of $\widehat{\gamma}$ even with biased estimation of $\widehat{\Delta}(j)$, we generate three binary covariates and a binary treatment using a multinomial random number generator with cell probabilities determined by the log-linear model:

$$\log\left(\mu_{ijk\ell}\right) = 2X_{i1} + X_{j2} - X_{k3} + T_\ell - X_{i1}X_{j2} - \lambda X_{i1}X_{k3} + \lambda X_{i1}T_\ell + (1+\lambda)X_{j2}X_{k3} + (1-\lambda)X_{j2}T_\ell - X_{k3}T_\ell, \quad (16)$$

where μ is the rate of sampling a subject with a particular combination of covariates and treatment. Again, the X_1 covariate defines the exclusion criterion (I) and X_3 is the potentially unmeasured variable (U). Note that the sufficient conditions (B1) and (B2) for unbiasedness of the generalizability bias estimator is satisfied whenever $\lambda = 0$. For simplicity, the other coefficients were chosen so that the size of the excluded subpopulation is approximately 70%. Also note that the log-linear model implies that the correct propensity model contains main effects for all three X covariates, but no interactions. The response variable is simulated according to regression model with constant variance of one and mean:

$$E\left[Y_i|\mathbf{X}_i, T_i\right] = X_{i1} + X_{j2} + X_{k3} + T_\ell + X_{i1}X_{j2} + X_{i1}T_\ell + X_{j2}X_{k3} + X_{j2}T_\ell + X_{i1}X_{j2}T_\ell + \lambda X_{i1}X_{j2}X_{k3}T_\ell. \quad (17)$$

Again for simplicity, each coefficient is set equal to one, except the coefficient for the interaction among all the coefficients and treatment. When this coefficient (λ) equals zero, this model satisfies sufficient condition (B3). In this simulation study, we only examine the case of unmeasured confounding in both the propensity and response estimation models.

4.2. Results

Each estimator is evaluated using the empirical bias, confidence interval coverage percentage, and confidence interval width. The Monte Carlo standard error for estimates of bias are less than 0.005 for all the estimators.

4.2.1. Unbiased $\widehat{\Delta}(j)$, Unbiased $\widehat{\gamma}$—We first evaluate the ability of each observational data subsample based estimator to estimate the SPATE in the eligible subpopulation, Δ(*I*). Figure 2 shows the bias of each estimator of Δ(*I*) for 75% exclusion, under the correct and incorrect model specifications. We see that, as expected, the simple estimator that does not control for any confounding performs poorly for estimating Δ(*I*). By design, our simulated observational data has confounding that the simple model does not adjust for. However, the other three estimators are approximately unbiased when the appropriate models are correct as indicated in Table 1. That is, $\widehat{\Delta}_{reg}$ performs well when the regression model is correctly specified, $\widehat{\Delta}_{IPW}$ performs well when the propensity model is correctly specified, and $\widehat{\Delta}_{DR}$ performs well when either the regression or propensity models are correctly specified. Each estimator becomes more robust to model mis-specification as the correlation increases between the covariates, since much of the information about the treatment choice in the omitted covariate is also carried by the remaining two covariates (and thus the model is in fact nearly correctly specified). For smaller rates of exclusion, the results (not shown) are similar but less pronounced.

We expect that when the estimates of SPATE perform well, the estimates of the generalizability bias will also perform well. Figure 3 shows this to be the case. For correctly specified models and large correlation (i.e., nearly correctly specified models), the estimates of generalizability bias perform well, having small bias and confidence interval coverage close to the nominal 95%. The exception is $\widehat{\Delta}_{DR}$ with a mis-specified regression model, whose coverage was slightly conservative due to the increased variance estimates. Nevertheless, of the estimators we considered, we prefer $\widehat{\Delta}_{DR}$, as it retains its robustness to either propensity or response model mis-specification, and produces confidence intervals with approximately nominal or conservative coverage whenever it is unbiased, while maintaining relatively narrow interval lengths. Again, the results (not shown) for smaller rates of exclusion are similar but less pronounced.

4.2.2. Biased $\widehat{\Delta}(j)$, unbiased $\widehat{\gamma}$—Based on the results of the multivariate normal covariates models, we expect our proposed estimators to perform well with respect to bias and confidence interval coverage whenever the response and propensity models are correctly specified. This is confirmed in Figure 4, which displays properties of the correctly specified estimators of SPATE(I) and generalizability bias as λ ranges from −1 to 1. As in the previous simulation study, the simple estimator makes no adjustments for confounding, and as such almost uniformly performs poorly (except in the coincidental case of λ = 0.48, where the bias in the eligible and ineligible samples are exactly equal according to our simulation model). In contrast, the estimators of SPATE(I) and generalizability bias that adjust for the confounding in the model are unbiased with approximately nominal or conservative confidence interval coverage, regardless of the value of λ.

It is often argued that the assumption of no unmeasured confounding is unreasonable in many real data analyses. Fortunately, our generalizability bias estimators are unbiased even with unmeasured confounding as long as the data meet the weaker sufficient conditions (B1)-(B3) described in Section 3.2. As noted in Section 4.1.2, when we omit the ‘unmeasured’ variable X_3, these conditions are met whenever λ = 0. The good performance of our estimators in this case is confirmed in Figure 5, which displays the properties of the estimators when X_3 is not included in either the propensity or the response models. Note from the upper left hand panel that all the estimators are all biased for SPATE(I), since they do not control confounding due to X_3. The CI coverage is correspondingly poor. However, when $\lambda \approx 0$, the covariate-adjusted estimators of generalizability bias are all approximately unbiased with approximately nominal or conservative confidence interval coverage. Thus, whenever the sufficient conditions approximately hold, our estimators perform well.

5. Example

To demonstrate our estimators in a realistic setting, we consider a randomized trial of two similar obstetric procedures – the use of forceps or vacuum cup in assisted vaginal delivery. Williams, et al. published the results of a prospective RCT comparing these procedures head to head [16]. This study considered several maternal outcomes, including length of hospital stay (LOS) and severe laceration. The inclusion criteria were predominantly the need for assisted vaginal delivery as determined by an attending physician. Among the exclusion criteria for this trial were fetal occiput transverse presentation, and the requirement that the mother be at least 18 years of age. We also apply the presentation exclusion, as the position could appropriately influence the choice of instrument. While not explicitly stated, we conjecture that the age requirement is related not to medical concerns, but the ability of the mother to legally consent to randomization, as 18 is the legal age of medical consent in most US states. We use our proposed methods to estimate the generalizability bias of a study of this type due to the age restriction.

Our data consist of a subset of the 2006 National Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality (AHRQ). NIS is a collection of electronic medical records for all patients discharged from a selection of hospitals across the United States. Among these were records for 56,300 women for whom assisted vaginal delivery was recorded without a diagnosis code for occiput transverse presentation. Each record includes LOS in days, and diagnosis of severe perinatal laceration (third or fourth degree). These data also recorded maternal traits, including age, obesity, economic status via payment type and median household income quartile for the patient's ZIP code, and urbanity of the patient's ZIP code; pregnancy traits, including high risk pregnancy diagnosis and weekend delivery indicator; and hospital information, including size, ownership, urbanity, teaching status, and region. We included all these variables as main effects in both the regression and propensity score models. Due to the large size of the database, we did not feel the need for further parsimony.

We extracted records for the 56,300 women for whom assisted vaginal delivery was recorded in 2006. We excluded 1381 records for which complete data on our covariates were not available, leaving 54,919 women for analysis. For the purpose of demonstrating our generalizability bias estimators, we assume that a complete case analysis that excludes only 2.5% of the records does not notably impact our inference. For a more thorough treatment, we recommend a more thoughtful approach to the missing data, such as multiple imputation [24]. Many of the selected covariates differ on average between treatment groups. For example, the proportion of assisted deliveries using forceps ranges from about 10% in the Northeastern and Western regions of the United States, to about 18% in the Midwestern and Southern regions.

We first examined the relationship between instrument type and LOS. Without adjusting for differences in covariates, we estimate that the use of vacuum shortens the average length of stay by approximately 0.05 days (see Table 2). This is consistent with, but smaller than, the 0.2 days estimated in the small clinical trial. We also attempted to adjust for potential confounding using regression, propensity weighted and doubly robust methods. For these methods, we found that available model diagnostics indicated no modeling concerns. In particular, we found good overlap in the propensity scores for the vacuum and forceps groups. For a more detailed discussion of diagnostics for propensity score methods see, e.g., [14, 15]. The propensity and doubly robust methods found significant average differences by procedure, but the regression method estimated a smaller and not statistically significant average difference. While we did attempt to adjust for measured confounding, we could not adjust for unmeasured and potentially confounding variables. Thus, while these estimates are fairly consistent across the estimators, they could all be over- or under-estimating the true PATE. Using experimental data may be preferable for drawing causal conclusions within the population recruited for the trial.

Next, we estimated the size of the generalizability bias in any RCT excluding women younger than 18 years of age. These estimators can be unbiased if the joint distribution of all the important variables in the population (including unmeasured variables) satisfies the sufficient conditions noted in Section 3.2. That is, there can be no unmeasured variable that is correlated with the adult status conditional on the covariates alone, or jointly conditional on the covariates and treatment. Because age is one of the covariates, it is unlikely that the adult status is significantly correlated with any unmeasured variable. Following the same reasoning, the adult status should also not interact with any unmeasured variables. Thus, the bias of $\hat{\gamma}$ is limited by the model mis-specification. Because the doubly robust estimator is robust to both the propensity and regression models, this is likely to be most accurate. Again, we found no reason for concern in the model diagnostics for both the adult and juvenile women. Table 2 displays all of the estimates as both raw and percent bias. Because none of the confidence intervals for γ excludes zero, we conclude that there is no significant bias in an RCT estimate of the effect of instrument choice on LOS due to excluding the young mothers.

We also examined the effect of instrument choice on severe laceration. As displayed in Table 3, we estimate that use of vacuum is statistically significantly associated with about 13% fewer severe lacerations. This is much larger than the insignificant 1% difference observed in the RCT, which indicates that our study may be failing to adjust for some important difference in the two groups. However, the same age-related arguments as for the LOS study apply here, leading us to believe our estimators for generalizability bias are themselves only biased by model mis-specification. Because the outcome of interest is binary, we omit the inappropriate regression model, and present only the naive and propensity-adjusted estimates in Table 3. While the simple method indicates no significant bias, the propensity-adjusted estimates all suggest that the effect estimate from the RCT would overstate the benefit of vacuum assistance over the use of forceps in the whole population. Thus, physicians may prefer to use caution when applying the RCT results to very young women. We do note that the magnitude of the bias estimates are only about 2% of the total effect. However, this is in part because only 5% of the population would have been ineligible for RCT participation due to age restrictions. If we supposed the exclusion rate were much higher, the bias would be proportionately larger. For example, the rightmost columns of Table 3 show an approximately 30% bias if the exclusion rate were in fact 75%, as is not uncommon in RCTs.

6. Discussion

We have proposed a class of estimators for generalizability bias, which results from using unbiased estimates of average treatment effect in a subpopulation to estimate the average treatment effect in the whole population. Through simulation, we have shown that our estimators perform well whenever the appropriate models for confounding are correctly specified. In this case, the regression and doubly-robust estimators, $\widehat{\gamma}_{reg}$ and $\widehat{\gamma}_{DR}$ have relatively narrow confidence intervals while maintaining the nominal coverage level. When the relevant models are mis-specified (the response model for the regression estimator, the propensity model for the IPW estimator, or both models for the doubly robust estimator), the estimators do not perform as well.

However, our second simulation study showed that our estimators can perform well even in the presence of uncontrolled confounding, as long as the bias is approximately the same in the two subpopulations. Heuristically, this balance in bias occurs whenever (1) any correlation between the unobserved variables and subpopulation membership can be explained by the observed covariates and treatment, and (2) there is no interaction between the unobserved variables and subpopulation membership in the response model. These sufficient conditions are described in detail in Section 3.2. In this case, the regression and DR estimators had close to nominal confidence interval coverage while maintaining a relatively narrow length.

Taken as a whole, our study indicates that the doubly robust estimator $\widehat{\gamma}_{DR}$ is reliable for estimating generalizability bias in a wide range of real data situations, since it is robust to mis-specification of either the response or the propensity score model, and still performs well even when both models are mis-specified as long as any important unmeasured variables are uncorrelated and not interacting with subpopulation membership in the manner of the sufficient assumptions described in Section 3.2.

While the theory behind the bias of our estimators presented in section 3 is broad, our simulation results regarding the confidence interval coverage are only specific to a small range of the possible data models. In particular, changing the distribution of the covariates to reflect skew or multimodality or including a nonlinear response model may lead to different results.

We have also shown the application of these methods to a realistic data situation. We used administrative records to study the bias of a study of assisted delivery. This type of data is becoming more readily available, and can be utilized to compare treatments in use at the same time. However, our methods do require data on both comparator treatments, as well as inclusion/exclusion criteria. Thus, this method is not particularly suited to studying the bias of a study of a new treatment not yet adopted by any part of the medical community. But it is very useful for new studies of comparative effectiveness using head-to-head comparisons of multiple treatments, as well as more general stage 4 monitoring of safety and effectiveness. In addition, it may be reasonable to consider estimated generalizability bias for an intervention in the same class as a newly proposed intervention to inform either study design or PATE estimation.

When studying the effect of treatment, the central question to most clinical research, the RCT is the design most often used and trusted. Because most clinical trials exclude a potentially large portion of the target population for scientific, ethical or cost reasons, it is unclear if the PATE in the sample population is generalizable to the population as a whole. It would be useful to analyze broadly representative observational data to assess the effect of treatment for a population. However, due to the possibility of confounding, observational data does not permit causal inference from naive analyses. We have shown that observational data can be useful in estimating sample selection error, even in the presence of confounding, and when the individual-level RCT data are not available. In fact, RCT data by definition do not directly contain any information about generalizability bias, and could introduce other sources of error if utilized in an estimate of generalizability bias. But, combining RCT and observational data could be very useful and important for providing greater external validity to the results of RCTs. In fact, our proposed estimators can be easily incorporated into simple cross design synthesis analyses [10] to improve estimation of PATE for population-wide regulation and the estimation of SPATE for subgroup decision-making.

Acknowledgments

We thank William Gardner and two reviewers for their helpful comments. The project described was supported in part by Award Number UL1RR025755 from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Research Resources or the National Institutes of Health.

References

1. Stirman SW, DeRubeis RJ, Crits-Christoph P, Brocly PE. Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. Journal of Consulting and Clinical Psychology. 2003; 71(6):963–972. DOI: 10.1037/0022-006X.71.6.963. [PubMed: 14622071]
2. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. Statistics in Medicine. 2008; 27:1801–1813. DOI: 10.1002/sim.3218. [PubMed: 18381709]
3. Humphreys K, Weisner C. Use of exclusion criteria in selecting research subjects and its effect on the generalizability of alcohol treatment outcome studies. American Journal of Psychiatry. 2000; 157:588–594. [PubMed: 10739418]
4. Zimmerman M, Chelminski I, Posternak MA. Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. Journal of Nervous and Mental Disease. 2004; 192:87–94. [PubMed: 14770052]
5. Fortin M, Dionne J, Phiho G, Gignac J, Almirall J, Lapointe L. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? Annals of Family Medicine. 2006; 4:104–108. DOI: 10.1370/afm.516. [PubMed: 16569712]
6. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, ShirtcliffFe P, Aldington S, Beasley R. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? Thorax. 2007; 62:219–223. DOI: 10.1136/thx.2006.066837. [PubMed: 17105779]
7. Rothwell P. External validity of randomised controlled trials: To whom do these results of this trial apply? Lancet. 2005; 365:82–93. DOI:10.1016/S0140-6736(04)17670-8. [PubMed: 15639683]
8. Altman DG. Poor-quality medical research: what can journals do? Journal of the American Medical Association. 2002; 287(21):2765–7. DOI: 10.1001/jama.287.21.2765. [PubMed: 12038906]
9. Eddy D, Hasselblad V, Shachter R. Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence. Harcourte Brace Jovanovich: Boston. 1992
10. Kaizar EE. Estimating treatment effect via simple cross design synthesis. Statistics in Medicine. 2011; 30:2986–3009. [PubMed: 21898521]
11. Greenland S. Multiple-bias modelling for analysis of observational data. Journal of the Royal Statistical Society: Series A. 2005; 168:267–306. DOI: 10.HH/j.1467-985X.2004.00349.x.
12. Marcus SM. Assessing non-constant bias with parallel randomized and nonrandomized clinical trials. Journal of Clinical Epidemiology. 1997; 50:823–828. DOI:10.1016/S0895-4356(97)00068-1. [PubMed: 9253394]
13. Wolpert RL, Mengersen KL. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. Statistical Science. 2004; 19:450–471. DOI: 10.1214/088342304000000350.
14. D'Agostino RB Jr. Propensity score methods for bias reduction for the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998; 17:2265–2281. DOI: 10.1002/0470023678.chlb. [PubMed: 9802183]
15. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research. 2011; 46:399–424. DOI: 10.1080/00273171.2011.568786. [PubMed: 21818162]

16. Williams MC, Knuppel RA, O'Brien WF, Weiss A, Kanarek KS. A randomized comparison of assisted vaginal delivery by obstetric forceps and polyethylene vacuum cup. Obstetric Gynecology. 1991; 78:789–794.

17. Hill JL. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics. 2011; 20:217–240. DOI: 10.1198/jcgs.2010.08162.

18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55. DOI: 10.1093/biomet/70.1.41.

19. Lunceford J, Davidian M. Stratification and weighting via the propensity score in estimation of casual treatment effects: a comparative study. Statistics in Medicine. 2004; 23:2937–2960. DOI: 10.1002/sim.l903. [PubMed: 15351954]

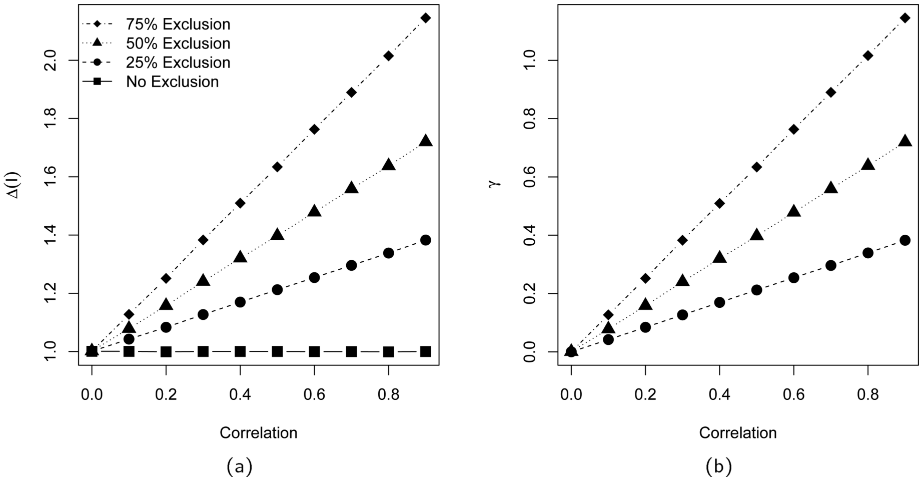
20. Kang J, Schafer J. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science. 2007; 22(4):523–539. DOI:10.1214/07-STS227.

21. Cochran, WG. Sampling Techniques. 3rd ed.. Wiley; New York: 1997. Section 6.18

22. Li L, Shen C, Wu AC, Li X. Propensity score-based sensitivity analysis method for uncontrolled confounding. American Journal of Epidemiology. 2011 advanced access, DOI: 10.1093/aje/kwr096.

23. Chaudhuri R, Livingston E, McMahon AD, Thomson L, Borland W, Thomson NC. Cigarette smoking impairs the therapeutic response to oral corticosteroids in chronic asthma. American Journal of Respiratory and Critical Care Medicine. 2003; 168:1308–1311. DOI: 10.1164/rccm.200304-503OC. [PubMed: 12893649]

24. Little, RJA.; Rubin, D. Statistical Analysis with Missing Data. 2nd ed.. Wiley; 2002.



(a)

(b)

Figure 1.

Population values for PATE (squares), SPATE(I) [panel (a)], and generalizability bias [panel (b)] as a function of covariate correlation. Values are displayed for simulations where 75% (diamonds), 50% (triangles), and 25% (circles) of the population is excluded from trial eligibility.

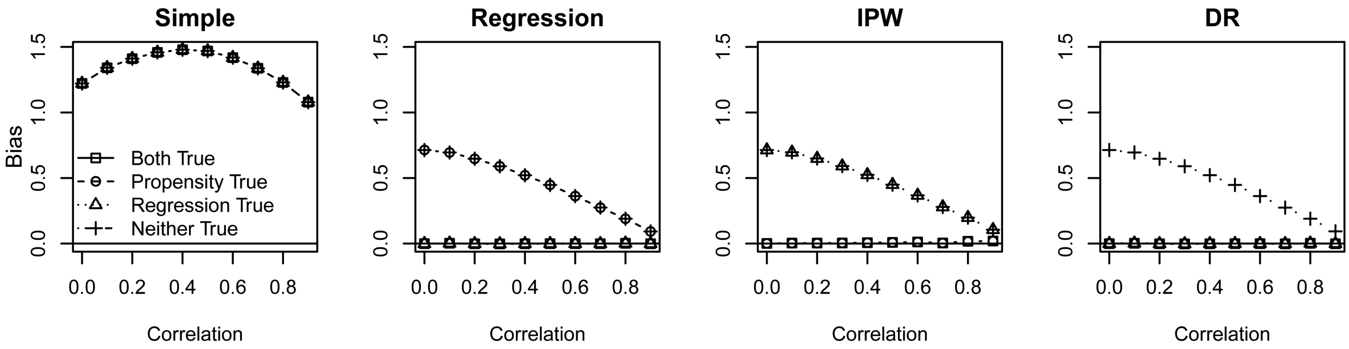


Figure 2.

Bias of using $\widehat{\Delta}(I)$ to estimate Δ as a function of covariate correlation, based on the first simulation study. Each panel displays results for estimates with different model misspecifications: correct propensity and response models (squares), correct propensity model and incorrect response model (circle), incorrect propensity model and correct response model (triangle), and incorrect propensity and response models (crosses).

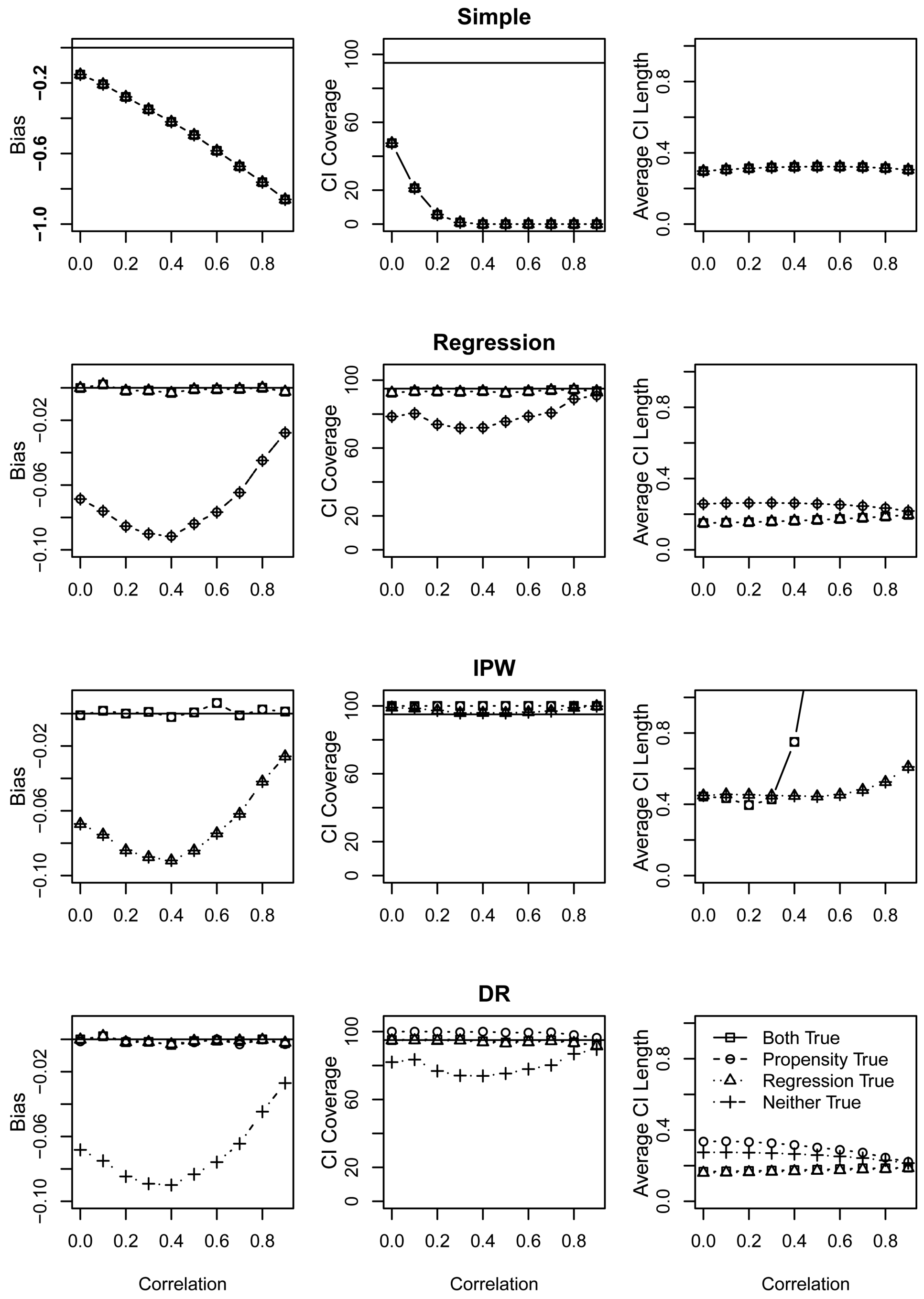


Figure 3.

Properties of $\hat{\gamma}$ as a function of covariate correlation based on the first simulation study. Each panel shows the properties for four model mis-specifications: correct propensity and response models (squares), correct propensity model and incorrect response model (circles), incorrect propensity model and correct response model (triangles), and incorrect propensity and response models (crosses). The panels in each row display the properties of $\hat{\gamma}$ based on the simple, regression, IPW, and DR SPATE estimators, respectively. The panels in the leftmost column display the bias of the estimator, where the solid horizontal line indicates no bias. Note that the vertical scale for the simple estimator is larger than the others to accommodate the much larger bias. The center column panels display the CI coverage, where the solid horizontal line indicates the nominal 95% level, and the rightmost column panels display the average CI length.

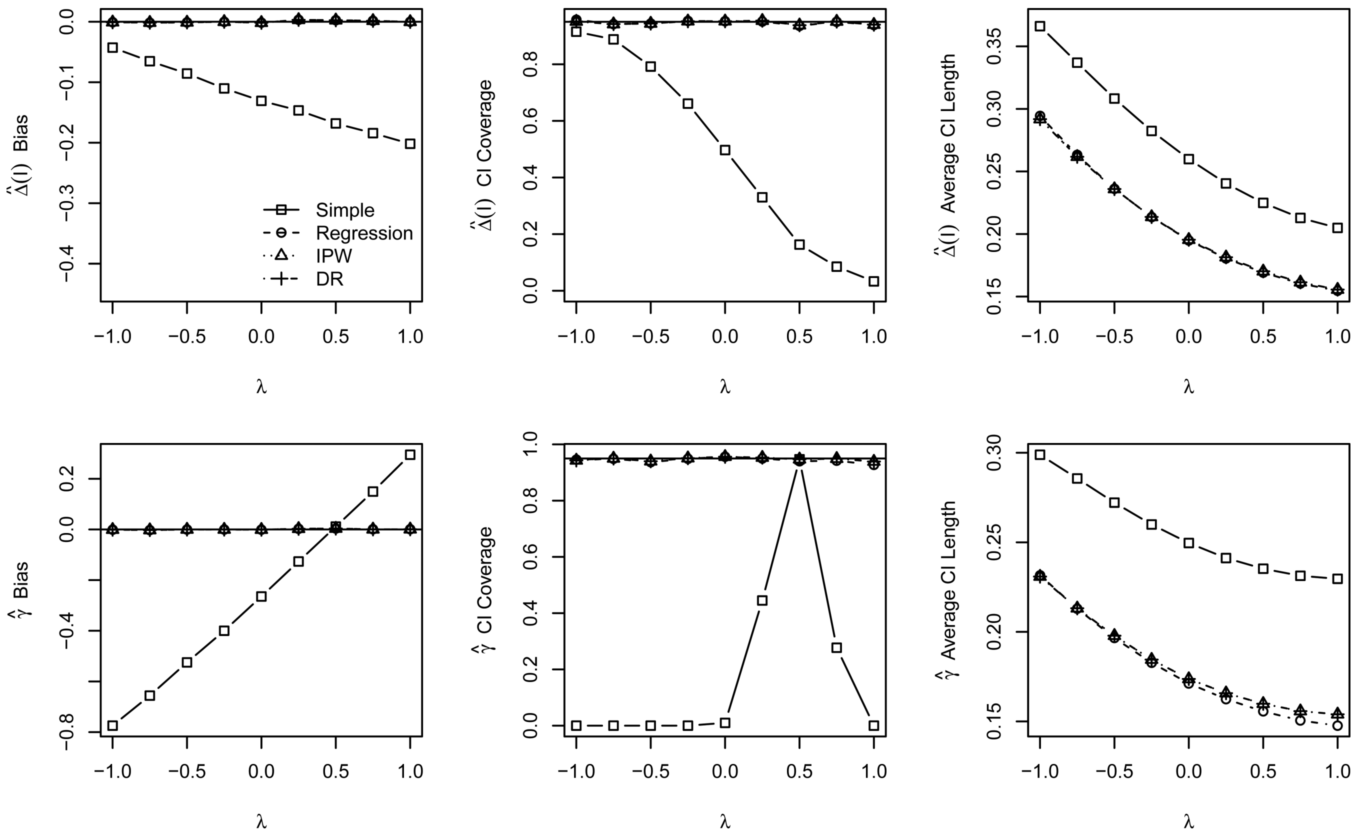


Figure 4.

Bias of using $\widehat{\Delta}\,(I)$ to estimate Δ (first row) and properties of $\widehat{\gamma}$ (second row) as a function of covariate correlation based on the second simulation study and correct model specifications. Each panel displays results for each of the four estimators: simple (squares), regression (circle), IPW (triangle), and DR (diamonds). The panels in the leftmost column display the bias of the estimator, where the solid horizontal line indicates no bias. The center column panels display the CI coverage, where the solid horizontal line indicates the nominal 95% level, and the rightmost column panels display the average CI length.

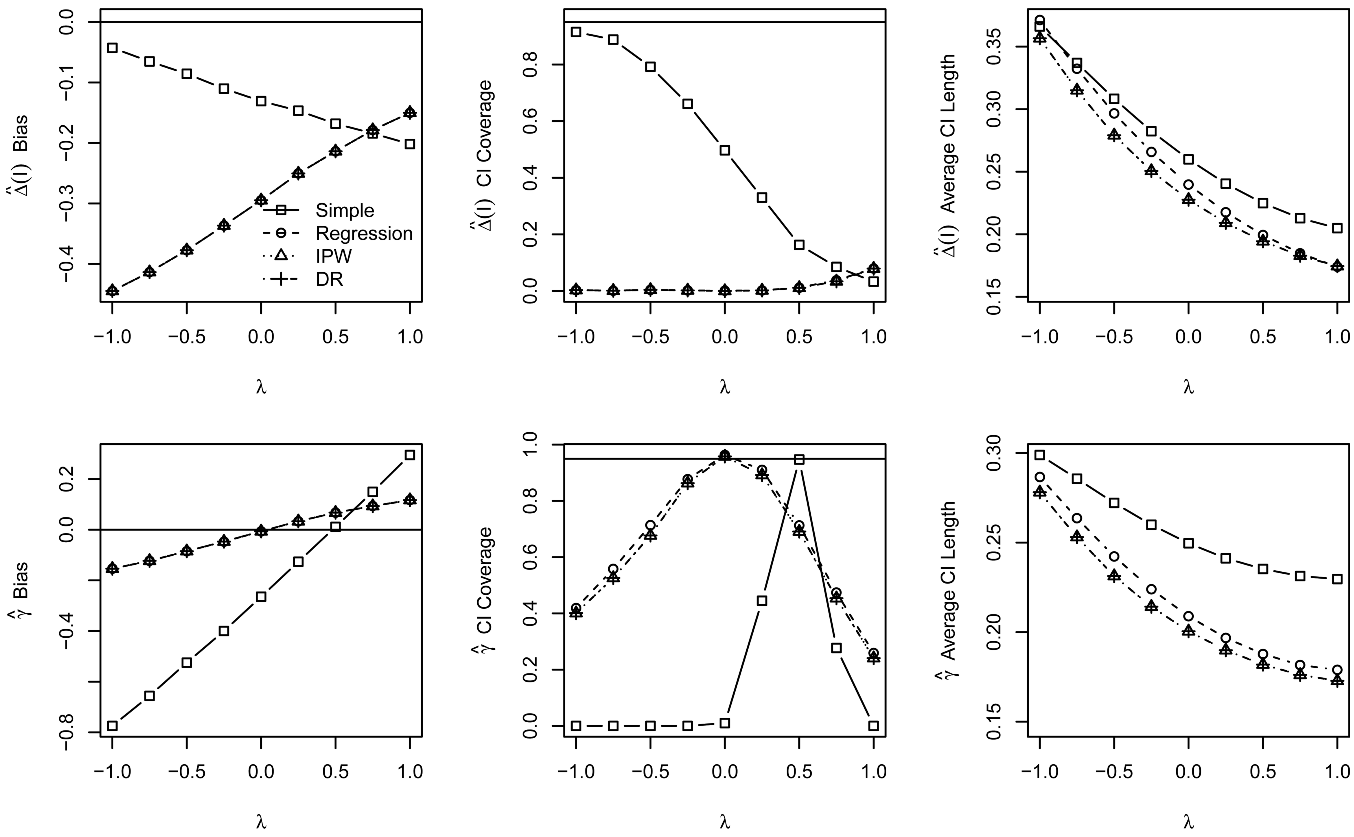


Figure 5.

Bias of using $\widehat{\Delta}\,(I)$ to estimate Δ (first row) and properties of $\widehat{\gamma}$ (second row) as a function of covariate correlation based on the second simulation study and incorrect propensity and respse model specifications. Each panel displays results for each of the four estimators: simple (squares), regression (circle), IPW (triangle), and DR (diamonds). The panels in the leftmost column display the bias of the estimator, where the solid horizontal line indicates no bias. The center column panels display the CI coverage, where the solid horizontal line indicates the nominal 95% level, and the rightmost column panels display the average CI length.

Table 1

Conditions for consistent estimation of Δ for each of four common estimators

A Estimator	**Sufficient Condition for Consistency**
Simple	No confounding
Regression	Correct response model
IPW	Correct propensity model
Doubly Robust	Correct response model or correct propensity model

Table 2

Estimates of the PATE and generalizability bias due to age exclusion in studies of the effect of assisted delivery instrument choice on LOS.

	$\hat{\Delta}$	$\hat{\gamma}$	**% Bias**
Simple	−0.05236 (−0.09164, −0.01308)	0.00325 (−0.00526, 0.01176)	−6.205
Regression	−0.03334 (−0.06839, 0.00172)	0.00139 (−0.00446, 0.00724)	−4.165
IPW	−0.04364 (−0.08169, −0.00560)	−0.00142 (−0.00802, 0.00518)	3.247
DR	−0.05037 (−0.08843, −0.01230)	−0.00123 (−0.00813, 0.00567)	2.442

Table 3

Estimates of the PATE and generalizability bias due to age exclusion in studies of the effect of assisted delivery instrument choice on severe laceration. The second and third column from the left display the estimates as described in Section 3.2, and the rightmost two columns display the estimates where SPATE is estimated from the data, but the proportion of the excluded population is artificially set to 0.75.

		$\hat{\pi}_E$ = 0.05		**π_E = 0.75**	
	$\hat{\Delta}$	$\hat{\gamma}$	**% Bias**	$\hat{\gamma}$	**% Bias**
Simple	−0.13042 (−0.14012, −0.12073)	−0.00169 (−0.00390, 0.00052)	1.297	−0.02511 (−0.05786, 0.00764)	19.250
IPW	−0.12805 (−0.13852, −0.11758)	−0.00259 (−0.00491, −0.00026)	2.021	−0.03842 (−0.07287, −0.00396)	30.000
DR	−0.12711 (−0.13758, −0.11664)	−0.00237 (−0.00469, −0.00004)	1.863	−0.03515 (−0.06964, −0.00066)	27.655