Received: 1 May 2020 | Revised: 11 January 2021 | Accepted: 5 February 2021

DOI: 10.1002/pst.2106

MAIN PAPER

WILEY

A comparison of reweighting estimators of average treatment effects in real world populations

Chen-Yen Lin[1] | Eloise Kaizar[2] | Douglas Faries[1] | Joseph Johnston[1]

[1]Eli Lilly and Company, Indianapolis, Indiana, USA

[2]Department of Statistics, Ohio State University, Columbus, Ohio, USA

Correspondence
Chen-Yen Lin, Eli Lilly and Company, Drop Code 1776, Lilly Corporate Center, Indianapolis, IN 46285-0001, USA.
Email: lin_chen_yen@lilly.com

Funding information
Eli Lilly and Company

Abstract

Regulatory agencies typically evaluate the efficacy and safety of new interventions and grant commercial approval based on randomized controlled trials (RCTs). Other major healthcare stakeholders, such as insurance companies and health technology assessment agencies, while basing initial access and reimbursement decisions on RCT results, are also keenly interested in whether results observed in idealized trial settings will translate into comparable outcomes in real world settings—that is, into so-called "real world" effectiveness. Unfortunately, evidence of real world effectiveness for new interventions is not available at the time of initial approval. To bridge this gap, statistical methods are available to extend the estimated treatment effect observed in a RCT to a target population. The generalization is done by weighting the subjects who participated in a RCT so that the weighted trial population resembles a target population. We evaluate a variety of alternative estimation and weight construction procedures using both simulations and a real world data example using two clinical trials of an investigational intervention for Alzheimer's disease. Our results suggest an optimal approach to estimation depends on the characteristics of source and target populations, including degree of selection bias and treatment effect heterogeneity.

KEYWORDS

entropy, external validity, propensity, weight estimation, weight trimming

1 | INTRODUCTION

Participants in randomized controlled trials (RCTs) are rarely chosen to resemble a simple, stratified or cluster random sample of any well-defined real world population. Even practical clinical trial designs rely on convenience samples or economic decisions that shift them away from true random sampling. In turn, if treatment effects systematically vary across individuals, raw treatment effect estimates reported from trials are not likely to be directly applicable to decisions confronting medical decision makers, with important consequences. For example, when evaluating new medications, a health plan may conclude that the patients enrolled in pivotal trials are too dissimilar to enrollees in its plan and decide to exclude or disadvantage the product from its formulary until "real world evidence" of effectiveness becomes available.

A growing number of studies have advocated reweighting methods to adjust trial-based estimators to be more relevant to specific target populations.[1-10] As a result of the importance of this topic to healthcare decision makers and stakeholders, use of new methods for generalizing evidence from RCTs was chosen as the focus for one of the five working packages for the Innovative Medicine Initiative Get Real Consortium.[11] Conceptually, the idea is to use statistical adjustments to estimate the average treatment effect had the trial had been conducted in a specific target population. While some of this literature starts to explore practical statistical performance of this type of estimator, there is still much to be learned about how such properties relate to practical choices such as exact estimator formulation and weight construction methodology. The aims of this paper are to begin to fill this knowledge gap using both simulated data and real data from two RCTs of a potential treatment for Alzheimer's disease.

Conceptually, we focus on the scenario in which a single RCT was conducted to evaluate the efficacy of a new intervention, hereafter termed the active treatment, versus control treatment (e.g., standard care or placebo). We assume a simple trial design so that efficacy (i.e., the sample average treatment effect) is reasonably estimated by the difference in mean outcomes between treatment groups. We define the target population with an observational dataset, which is possibly a subset of a larger database. The observational dataset is considered to resemble a simple random sample from the target population, for which a census is a special case. In contrast, we assume that the RCT participants do not resemble a simple random sample from the target population. Borrowing ideas from survey sampling, generalization or transportation reweighting estimators assign a weight to each RCT participant so that the weighted distribution of participant characteristics resembles that of the target population. Weighting can, in theory, be used to overcome selection biases introduced by standard recruitment practices. For example, it is well known that African Americans are typically under-represented in United States (U.S.) trials, even when race is not an inclusion/exclusion criterion.[12] In the RCT context, selection bias often refers to biases that result from differences in subject characteristics between treatment arms, but it is also defined to encompass differences across other subject-specific conditions.[13] Throughout this manuscript, we use the term selection bias to indicate biases that result from imbalance in subject characteristics between trial participants and the target population.

However, if there is no RCT participant that represents some characteristic of the target population that is deemed potentially important to the treatment effect (e.g., the RCT cohort includes only patients with mild symptoms, but the target population includes patients with mild through severe symptoms), then no set of RCT weights could lead to the target distribution. This phenomenon is akin to violation of the positivity assumption of causal inference,[5,14,15] and the reweighting-only estimation may be hopelessly inconsistent. One approach to overcoming this lack of positivity is to use another source of information about treatment effect for those not represented by anyone enrolled in the RCT, for example, via cross design synthesis.[16-20] Because such extra data are typically not available during drug development, we are forced to conceptually avoid this difficulty by specifically excluding individuals that could not be enrolled in the RCT from the target population. That is, we require that the target population be some subset of the collection of all trial-eligible individuals. To this end, we apply as many of the RCT exclusion criteria likely to be relevant to treatment effect as practical (often originally imposed to enhance safety, thrift, or statistical power) to the observational database, and hence our target population. In this way, functions of weighted averages of the RCT participant outcomes can potentially reflect the average treatment effect in the target population.

We study two types of estimators. One is of the same form as inverse propensity weight (IPW) estimators, which have long been used to adjust for selection effects in analysis of probability sample surveys, and more recently in causal inference from observational data.[21] Our second type of estimator combines IPWs with a parametric model of the outcome. This general approach also has a rich history in survey sampling[22] and more recently in causal inference[23] and extending inferences.[8] In the latter two cases, methods that rely on both weights and models are often termed “robust” or “doubly robust.” In the probability sampling framework, IPWs are the inverse of the probability that an individual is chosen to participate in the study. As such, the sampled individuals “represent” a number of individuals in the same population from which they were randomly sampled that is proportional or equal to their weight. Thus, ideally, the weighted distribution of any variable in the sample resembles its distribution across the whole population. In extending this idea to the generalization application, we typically wish to preserve this characteristic: key attributes of weighted sample distributions resemble those attributes of distributions in a target population. An algorithmic approach can be used to specify weights to achieve this same distributional features matching without explicitly estimating the probability of study participation.

Weighted representations of particular populations arise in the causal inference setting as well. Weights can be assigned separately to those who received the control and active treatments, so that each of the two group-specific distributions of characteristics resemble the single marginal (combined treatment group) distribution of these characteristics. Thus, weighting may reduce possible confounding by observed characteristics whose distributions differ across the two treatment groups. By comparing the group-specific IPW-weighted distributions of the outcome of interest, we can learn about the treatment effect in the entire combined study sample. For example, the difference in the weighted mean in the control group and the weighted mean in the active treatment group may be a good estimate of the population average treatment effect, where we define the population of interest to be the entire set of study participants (or a population related to it by simple random sampling). The validity of the comparison between weighted distributions rests on treating the two treatment groups as if they were each separate random (probability) samples from this single target population.

For our trial generalization work, we similarly weight the participants of a RCT to represent a single target population. This approach differs from traditional causal inference in two key ways: (1) the target population is defined separately from the trial participants, and (2) the treatment is randomly assigned to participants. The former indicates that we need external information to define the target population, similar to how a probability sampling survey design defines the population. The latter suggests we do not need to distinguish between the two arms in the model of selection, since randomization is expected to balance characteristics across the two trial arms, particularly in large scale late phase clinical development programs, and thus lessen concerns about confounding.

In practice, we typically do not know each individual's probability of trial participation, and thus their respective weights—even for designs that invite participation via probability sampling. The weights must be estimated. Estimation methods can be roughly put into two categories: (1) a model-based approach that inverts a predicted probability of selection and (2) mathematical algorithms that calculate weights that directly minimize the imbalance in characteristics. While the merits of these respective methods have been investigated elsewhere,[24] we are not aware of any other exploration of the practical implications of the choice of estimator in the generalizability context. We focus on one implementation of each approach to study the practical implications of this choice on the statistical properties of reweighted treatment effect estimators.

For the model-based approach to weight calculation, we use logistic regression, which is arguably the most commonly implemented method for estimating propensity scores. In principle, we could use other soft classifiers, including but not limited to bagging, boosting,[25] and random forests.[26] For direct weight calculation, we use entropy balancing, although a number of similar approaches have also been investigated.[24,27]

Returning to classical survey sampling, researchers have long known that the variance of weighted estimators can become quite large whenever the weights themselves have a large variance.[28] One intuitive explanation is that if a small number of individuals account for most of the weight, the effective sample size (and thus the precision) for the weighted estimator is also small. Trimming the weights to some maximum value (and redistributing the trimmed weight to the other study participants to maintain their representative interpretation) reduces the variability of the weights, and in turn reduces the variability of the weighted treatment effect estimator. Although weight trimming also introduces bias when the weight calculation/estimation is correct, this cost can pay for itself with a reduced mean squared error.[29,30]

Weight trimming should similarly reduce the variance of trial generalization estimators of average treatment effects. However, the relative impact on bias and the practical implications of this variance-bias tradeoff have not previously been studied in the generalization context. Several weight trimming methods have been proposed,[30,31] but we focus on two exemplars—a prespecified inflation factor cutoff and a prespecified percentage trim.

The paper is organized as follows. In Section 2, we formally define the population estimand, estimators, and methodology for constructing and trimming the weights. In Section 3, we describe and summarize two separate simulation studies focused on the two approaches to weight estimation and weight trimming and evaluate the impact of alternative approaches in scenarios involving varying degrees of selection bias and heterogeneity of treatment effect. We demonstrate the implementation of the reweighting methods and trimming in a real case study in Section 4, and discuss our recommendations in Section 5.

2 | STATISTICAL METHODS

We begin by clearly defining our population estimand, the target population average treatment effect (TATE), before turning to weight calculation and trimming, and finally introducing a collection of weighted estimators.

2.1 | Estimand: Target population average treatment effect

We use Rubin's causal model (RCM)[32] as a building framework to define the estimand, TATE. In the RCM, each individual has potential outcomes, $Y_i(1)$, and $Y_i(0)$, which would be the outcomes had the i-th subject received active and control treatment, respectively. The treatment effect for this individual is the difference in their potential outcomes, $TE_i = Y_i(1) - Y_i(0)$. It follows that the TATE is the average treatment effect across the individuals in that population:

$$\text{TATE} = \frac{1}{n_{\text{target}}} \sum_{i \in \Gamma_{\text{target}}} TE_i = \frac{1}{n_{\text{target}}} \sum_{i \in \Gamma_{\text{target}}} (Y_i(1) - Y_i(0)), \tag{1}$$

where n_{target} is the number of individuals in the target population and Γ_{target} is the collection of indices for individuals in the target population. Equation (1) serves as a conceptual quantity of interest that cannot be measured in practice, since the treatment effects for those in the target population are not measurable for many reasons. One obvious reason is that a subject cannot typically receive both active and control treatments. Secondly, target population members may not have access to the active treatment (as in the case of a drug that is unavailable to the target population). Third, the outcome measure is often not available in data describing the target population. For the remainder of this paper, we assume these three practical limitations. In addition, the specific individuals in the target population and their attributes, $\boldsymbol{X}_i$, may change over time and so measured characteristics resemble a random sample from some common distribution of interest. Thus, our well-defined estimand is the expected value of the TATE over both the distributions of potential outcomes and of relevant characteristics of individuals in the population:

$$E(\text{TATE}) = E\left(\frac{1}{n_{\text{target}}} \sum_{i \in \Gamma_{\text{target}}} Y_i(1) - Y_i(0)\right) = E_{X_i}(E(Y_i(1) - Y_i(0)|\boldsymbol{X_i})) = E_{X_i}(E(Y_i(1)|\boldsymbol{X_i})) - E_{X_i}(E(Y_i(0)|\boldsymbol{X_i})), \tag{2}$$

where $\boldsymbol{X}_i^T = (x_{i1} \ldots x_{ip})$ is a p-dimensional vector of characteristics, and subscript i indexes an individual who is randomly selected from the target population (with equal probability). An alternative interpretation of this quantity is the center of the distribution of TATE across many similar populations, such as an insurance pool on a randomly selected day.

To ensure that both TATE and E(TATE) are identifiable, we follow the generalization of typical RCM assumptions for the treatment assignment and outcome distributions, including consistency (i.e., the stable unit treatment value assumption) and positivity (the probability of receiving either treatment is theoretically strictly positive).[5,14]

2.2 | Weight calculation

As mentioned in the introduction and discussed in more detail in Section 2.3, our proposed generalization reweighting estimators are based on individually weighting each RCT participant. We derive the subject weights using two approaches: a model-based approach and an algorithmic balancing approach. For the weighting approach to work, it must be possible for every member of the target population to be statistically represented by RCT participants. This condition is closely related to the positivity assumption in the previous section. By putting the assumption in the generalization context, relevant exclusion criteria implemented in the trial must also be applied to the target population so that everyone in the target population would have a positive chance to be enrolled in a trial. For example, if the RCT explicitly only recruits patients with mild-to-moderate Alzheimer's disease, then we must exclude anyone with severe disease from the target population. Those with severe disease would have had zero probability of being enrolled in to the RCT. Unlike the traditional causal inference setting, we do not require the reverse to be true. That is, RCT participants may have zero probability of being included in the target population.

Under the model-based approach, we weight each subject by the inverse of their propensity score. Here we use the term propensity score in a different context from its common definition in the causal inference literature, based on a "propensity to choose treatment." In this paper, we instead use the "propensity to enroll in a RCT," that is, the conditional probability of an individual included in either a RCT or observational database being a RCT participant. In this context, the generalization propensity score is defined as $\Pr(S_i = 1| \boldsymbol{X}_i)$, where binary variable $S_i = 1$ and $S_i = 0$ indicates that the i-th individual is included in a trial and part of the observational database, respectively.

We define the model-based subject weight to be

$$\tilde{W}_i = \frac{1 - Pr(S_i = 1|\boldsymbol{X}_i)}{Pr(S_i = 1|\boldsymbol{X}_i)} I_{\Gamma_{\text{trial}}}(i), \tag{3}$$

where I is the indicator function. The subject weight in Equation (3) is reminiscent of the weight commonly used in the IPW estimator frequently employed for casual inference. The odds formulation implies that each individual's weight equals the number of individuals he or she "represents" in the target population. Note that only trial participants have positive weight, since they are the only individuals who may contribute to the weighted estimator.

The model-based generalization propensity score must be estimated by leveraging subject-level data from the combined RCT data and observational database. For our implementation, we follow the most common practice by assuming the conditional probability of RCT inclusion, membership variable S_i's are conditionally independent Bernoulli random variables and the conditional probability $\Pr(S_i = 1|\, \boldsymbol{X}_i)$ can be characterized using a logistic model of the form:

$$\log\left(\frac{\Pr(S_i=1|\boldsymbol{X_i})}{1-\Pr(S_i=1|\boldsymbol{X_i})}\right)=\alpha_0+\boldsymbol{\alpha^T X_i}, \tag{4}$$

where $(\alpha_0\ \boldsymbol{\alpha^T}) = (\alpha_0\ \alpha_1 \cdot \cdot \cdot \alpha_p)$ is a vector of unknown parameters that can be estimated using maximum likelihood estimation. In the logistic model setting, the estimated propensity-based weights can be written explicitly, $\tilde{W}_i = 1/\left(e^{\hat{\alpha}_0+\hat{\alpha}^T X_i}\right), i \in \Gamma_{\text{trial}}$. We use iteratively reweighted least squares via the glm R function[33] to estimate the model parameters.

In addition to model-based weights, we propose using an algorithmic approach to calculate individual weights directly without considering generalization propensities. In particular, we implement entropy balancing.[34] The conceptual framework of this approach is straightforward: adjust patient weights from the RCT such that the weighted moments of the RCT participants exactly match (or balance) those in the target population. Unlike propensity weights, entropy weights do not have a parametric representation nor an analytical form, but are calculated numerically by optimizing the objective function:

$$\min_{\tilde{W}} -1 \sum_{i\in\Gamma_{\text{trial}}} \tilde{W}_i \log\frac{\tilde{W}_i}{q_i} \tag{5}$$

such that $\left(\sum\limits_{i\in\Gamma_{\text{trial}}} \tilde{W}_i\right)^{-1} \sum\limits_{i\in\Gamma_{\text{trial}}} \tilde{W}_i X_{ij}^k = n_{\text{target}}^{-1} \sum\limits_{i\in\Gamma_{\text{Target}}} X_{ij}^k,\ \ \text{for}\, j=1,...,p, \tilde{W}_i = 0\, \text{for}\, i\in\varGamma_{\text{Target}}, \text{and}\ k\in K\subset\mathbb{N}^+$, where q_i is the base weight and K is the set of natural numbers that identify the first k moments to be matched. The minimand in Equation (5) is the so-called entropy divergence,[35] hence giving rise to the name entropy balancing. The constraints ensure that weighted moments for the trial population exactly equal those of the target population. Unless prior knowledge about the RCT recruitment plan would suggest otherwise, we propose setting an equal base weight at $1/n_{\text{trial}}$, the inverse of the trial size. We use the R package ebal[36] to implement the simplest entropy balance where $K = 1$ so that only the marginal means of $\boldsymbol{X}$ are balanced.

Regardless of whether we use model- or algorithm-based methods, we expect the weights to in some sense "balance" the characteristics $\boldsymbol{X}$ between the trial and target population by adjusting the empirical distribution of $\boldsymbol{X}$ among the RCT participants to resemble its distribution in the target population. The model-based propensity method has long been popular in the causal inference literature. It is particularly attractive because of the straightforward sampling probability interpretation, practitioner familiarity with logistic regression, and the ease with which we can explore potential causes for selection bias (by interpreting the model parameters $\boldsymbol{\alpha}$) and degree of imbalance (by comparing weight distributions across the RCT and target samples). However, algorithm-based balancing methods are gaining a stronger foothold because they result in an exact distributional balance with respect to the predefined characteristic. For the generalization setting, this is particularly simple because only summary data, such as the first few moments of the target population, must be specified in Equation (5). On the contrary, model-based propensity methods, such as logistic model, typically require subject-level data in both RCT and target populations.

Irrespective of which method we use to obtain initial weights, as a final step, we rescale each collection of weights such that the sum of weights in the trial population is equal to the size of the trial population. The final weight has the form:

$$W_i = \frac{\tilde{W}_i}{\frac{1}{n_{\text{trial}}}\sum_{j\in\Gamma_{\text{trial}}}\tilde{W}_j}, i\in\Gamma_{\text{trial}}. \tag{6}$$

The rescaling gives an intuitive interpretation. W_i loosely represents the number of people similar to trial participant i that should have been included in the trial so that, when taken together, the trial participants would naturally represent the target population.

In addition to these "final" weights, we consider post-processing trimming procedures in an effort to alleviate large variances for our E(TATE) estimates. Two trimming procedures are considered: a prespecified inflation factor cutoff, and a prespecified percentage trim. The first approach explicitly sets the upper bound of the final weight of any individual at a constant C. That is, we adjust the weight of any trial participant who was representing more than C times the number of individuals that they would have under simple random sampling. The trimmed weights are:

$$W_i^*(C) = \left(\frac{\min(W_i, C)}{\frac{1}{n_{\text{trial}}} \sum_{j \in \Gamma_{\text{trial}}} \min\left(W_j, C\right)} \right). \tag{7}$$

Despite the many systematic methods for choosing the constant C presented in the survey sampling literature,[31] many real-world survey analysts continue to choose the cutoff for unacceptably large weights, C, in an ad-hoc manner. For example, the National Assessment of Educational Progress (NAEP) survey uses cutoffs similar to C = 3.5 or 4.5[37]; the National Health and Nutrition Examination Survey (NHANES) uses cutoffs close to C = 3.[38] In this type of ad hoc approach where C is determined prior to data collection, if no final weights are relatively large, then the weights are not trimmed at all. Other approaches attempt to identify unusually large weights. For example, the National Immunization Survey (NIS) trims weights that are more than three times the interquartile range larger than the median weight.[39] A similar simple method is to simply trim weights that exceed some empirical percentile so that $C = W_{[\lfloor \mathbb{p} * n_{\text{trial}} \rfloor]}$, where $\mathbb{p}$ is the pre-specified percentile, square brackets indicate order statistics and $\lfloor \cdot \rfloor$ is the floor function. The percentile approach guarantees some predetermined number of participants have trimmed weights. It has been explored in a similar weight-based causal inference context.[15,30]

2.3 | Estimators

We propose three possible estimators of average treatment effect, each of which relies on computed weights. As noted previously, we use a single set of weights to overcome the limited external validity of the RCT and rely on the strong internal validity imparted by the random assignment to treatment to make causal inference. As a result, no further adjustment is made to address imbalance between treatments even though deviations from an ideal trial could be present and might undermine this internal validity, as discussed in the Introduction. Nonetheless, we recommend in practice checking balance of key covariates across treatment groups after reweighting.

We consider three different weight-based estimators of E(TATE) that follow one of two general approaches to estimation. First, motivated by the simple IPW estimator in causal inference, we propose a nonparametric approach by separately estimating each of the counterfactual averages in the final line of Equation (2) and then subtracting:

$$E\widehat{(\text{TATE})}_{np} = \frac{1}{\sum\limits_{i \in \Gamma_{\text{trial}}} W_i Z_i} \sum_{i \in \Gamma_{\text{trial}}} Y_i W_i Z_i - \frac{1}{\sum\limits_{i \in \Gamma_{\text{trial}}} W_i (1 - Z_i)} \sum_{i \in \Gamma_{\text{trial}}} Y_i W_i (1 - Z_i), \tag{8}$$

where Y_i is the outcome variable measured in the RCT and Z_i is an indicator of active treatment. This procedure mimics the simplest analysis of a simple RCT (without stratification) for which participants are a simple random sample from the target population. Under proper regularity conditions, the nonparametric estimator is asymptotically consistent. The proof follows the same form as the usual IPW estimator,[14] and the simulation study presented in Section 3 provides further evidence for such consistency.

In our second approach, we consider two estimators that mimic a slightly more complex analysis of a simple trial. As a result of its ease of interpretation and potential for smaller standard errors, we assume a parametric form for the conditional expectations in the last line of Equation (2). A parametric form also lays the groundwork for generalizing from more complex trial designs. We posit a linear regression model to characterize the relation between the outcome and treatment, baseline characteristics and their interactions:

$$Y_i = \mu + \gamma Z_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\theta}^T \boldsymbol{X}_i Z_i + \varepsilon_i, \varepsilon_i \sim \left(0, \sigma^2\right) \tag{9}$$

where μ, γ, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown parameters, and ε_i's are independently and identically distributed (*iid*) random errors with mean zero and common variance $\sigma^2 < \infty$. This linear model implies a concise representation of E(TATE), which depends on the model parameters and characteristics in the target population:

$$E(\text{TATE}) = \gamma + \boldsymbol{\theta}^T E_{\boldsymbol{X}|S=0}(\boldsymbol{X}) \tag{10}$$

In this parametric representation, the treatment effect heterogeneity is explicitly quantified through $\boldsymbol{\theta}$.

The model parameters can be estimated from the trial data using a weighted least squares (WLS) approach so that the distribution of the covariates $\boldsymbol{X}$ resembles that in the target population. A weighted least squares estimator also mitigates any effects of model mis-specification, similarly to design-based regression in the survey setting.[40,41] We complete plug-in estimators of E(TATE) by estimating the conditional expectation of $\boldsymbol{X}$ in the target population via one of two methods. We use either a weighted average from trial participants, or a sample average from the target population, respectively denoted as.

$$\begin{aligned} E\widehat{(\text{TATE})}_{p-\text{trial}} &= \hat{\gamma} + \hat{\boldsymbol{\theta}}^T \left(\frac{1}{n_{\text{trial}}} \sum\nolimits_{i \in \Gamma_{\text{trial}}} W_i \boldsymbol{X}_i \right) \\ E\widehat{(\text{TATE})}_{p-\text{target}} &= \hat{\gamma} + \hat{\boldsymbol{\theta}}^T \left(\frac{1}{n_{\text{target}}} \sum\nolimits_{i \in \Gamma_{\text{target}}} \boldsymbol{X}_{\boldsymbol{i}} \right) \end{aligned} \tag{11}$$

One noteworthy point is that these two estimators coincide for un-trimmed entropy balanced weights. However, they may diverge for propensity-based weights, and the difference may be large for small sample sizes or if the propensity model is mis-specified. Extra caution should be taken when these two estimates do not agree, and diagnostic procedures may help identify potential model misspecification.

2.4 | Variance estimation

The analytical form of the estimators is presented above, but the complex correlation makes it prohibitive to render a closed form variance function. Some large sample approximations are available for the variance of some propensity balanced estimators when the weights are based on logistic regression.[42] However, to our knowledge, such approximations are not derived more broadly for all the estimators described in Section 2.3. Instead, we rely on bootstrap methods to estimate the variance of our estimators. In particular, we create multiple pseudo datasets by randomly sampling with replacement subjects from both the RCT and target populations and use the pseudo datasets to carry out the estimation. This nonparametric bootstrap approach is repeated multiple times, and we use the empirical variance across the replicates as a variance estimate. In addition to variance estimation, nonparametric bootstrap also allows us to construct two types of confidence intervals. One is based on normal approximation by plus/minus a normal quantile times the square root of the variance estimate; the other is based on the equal-tailed lower and upper empirical quantiles.

3 | SIMULATION STUDY

In all, we have described two methods of calculating weights, two methods of trimming weights, and three different estimators of E(TATE). Each method has a theoretical or practical advantage. We use simulation to explore the practical impact of these choices in a range of realistic settings.

3.1 | Simulation methods

3.1.1 | Data generating mechanisms

The data generation consists of two components: a selection data generating mechanisms (DGM) and an outcome DGM. The selection mechanism determines the subjects' trial membership (trial participants versus target population) based on their characteristics; the outcome mechanism governs the relationship between the outcome of interest and subject characteristics. In this demonstration, we use a logistic model and a linear model as noted in Equations (4) and (9) for the selection and outcome mechanisms, respectively.

In the selection mechanism, the logistic model parameters, $\boldsymbol{\alpha}$, are chosen such that we expect 20% of the subjects will be assigned to RCT population (i.e., the marginal odds of trial membership is 1:4). Within this framework, we consider three simulation scenarios by varying the overlap in the characteristics between trial and target population via the magnitude of model parameters (α's, including setting some $\alpha_j = 0$). We will elaborate on how to quantify the degree of overlap in the next sub-section. The number and magnitude of characteristics' effects are shown in Table 1.

Treatment assignment and outcome variables are generated per specifications of the outcome mechanism for the trial participants only. For the treatment assignment, we randomly assign half of the trial participants to the active treatment. For the outcome, we follow the linear model structure given in Equation (9), where the errors follow a standard normal distribution. The overall mean μ and main characteristic effects $\boldsymbol{\beta}$ are not contributing to the E(TATE), thus we arbitrarily set all these parameters equal to zero. Similarly, the size of the baseline treatment effect is not of primary interest (although it does affect absolute estimator properties, such as relative bias); we set $\gamma = 2\sigma^2 = 2$.

We consider three different patterns of heterogeneous treatment effect, $\boldsymbol{\theta}$, as defined by the outcome mechanism parameters listed in the four scenarios in Table 2. (Note that Scenarios C and D have the same overall pattern of heterogeneous treatment effect, but with covariate numbering schemes as explained below). First, we consider a "moderate narrow" treatment effect heterogeneity, where four of the p characteristics moderately influence treatment effect and

TABLE 1 Parameters used in simulating selection model

Selection mechanism parameter or summary	Scenario		
	1	2	3
	Small	Uneven broad	Strong focused
α_0	−3.4	−11.1	−5.15
α_j	$\alpha_{1,2,3,4} = 1$ $\alpha_{5-16} = 0$	$\alpha_{1,2,3,4} = 3$ $\alpha_{5-16} = 0.5$	$\alpha_1 = 6$ $\alpha_{2-16} = 0$
Δ_p	0.05	0.31	0.28
$\mathcal{B}_T$	0.96	0.76	0.76

TABLE 2 Parameters used in simulating outcome model

Outcome mechanism parameter or summary	Scenario			
	A	B	C	D
	Moderate narrow	Moderate broad	Strong focused	Strong focused
μ, β_{1-16}	0	0	0	0
γ	2	2	2	2
θ_j	$\theta_{1,2,3,4} = 1$ $\theta_{5-16} = 0$	$\theta_{1-16} = 0.25$	$\theta_1 = 5$ $\theta_{2-16} = 0$	$\theta_2 = 5$ $\theta_1 = \theta_{3-15} = 0$
σ^2	1	1	1	1
R^2	0.46	0.41	0.67	0.67
CV	14%	7%	32%	32%

Abbreviation: CV, coefficient of variation.

the other have no association with the outcome. Second, we suppose a "moderate broad" heterogeneity where all p characteristics affect the treatment effect, but only slightly. Finally, we consider a "strong focused" heterogeneity where heterogeneity is driven by a single very influential characteristic. In this last case, we implement simulations where this single covariate is the same as for the selection model (the first covariate for Scenario C) and where it is different (the second covariate for Scenario D).

For all simulations, we set the number of characteristics, $\boldsymbol{X}$, at $p = 16$, and generate these independently for each subject from mutually independent Uniform(0,1) distributions. We run 500 repetitions for each combination of scenarios in selection and outcome mechanisms.

We focus on a large sample situation by setting the sum of trial and target sample sizes, $|\Gamma_{\text{trial}}| + |\Gamma_{\text{target}}|$, to 3000 (a count similar in size to trials and observational data that are, in our experience, typical for Alzheimer's disease research). For each simulated dataset, we estimate E(TATE) using all three versions of the proposed estimator in Equations (8) and (11) without weight trimming and with trimming. Because our logistic regression propensity estimation models are defined correctly, we anticipate any weight trimming will introduce some bias, and thus propose relatively conservative cutoff constants via the ad-hoc values $C = 4$ and $C = W_{[\lfloor 0.99 * n_{\text{trial}} \rfloor]}$, the 99th empirical percentile. The former value is consistent with the cutoff used in some large sample surveys; the latter is consistent with the most conservative threshold considered by Cole and Hernan.[15] The standard error of the proposed estimator is approximated using nonparametric bootstrap with 200 resamples, as described in Section 2.4.

3.1.2 | Measures of data generating mechanism and model performance

We first introduce a few useful metrics to quantify (1) the degree of overlap for the selection mechanism and (2) the magnitude of signal-to-noise for the outcome mechanism, and then explain (3) how to evaluate the performances of the different reweighting estimators.

To begin with the selection mechanism, excessive distributional imbalance between the trial and target populations theoretically suggests that E(TATE) estimates have large error and, as such, are of limited practical use. We quantify this imbalance with two previously proposed measures. First, Stuart et al. proposed looking at the difference in mean propensity scores, denoted as Δ_p. Here, larger scores indicate less overlap; while the effect of overlap has not yet been extensively studied in the generalizability context, differences greater than 0.25 or 0.1 standard deviations raise concerns about estimator stability and sensitivity to the propensity model specification in the usual causal inference context.[43] Second, Tipton quantified the affinity in propensity score distributions between the groups, denoted as $\mathcal{B}_T$, by the notion of the Bhattacharyya Coefficient.[44,45] Tipton suggested estimates of $\mathcal{B}_T < 0.5$ to indicate that the estimate E (TATE) may suffer from mean squared error (MSE) large enough to provide little useful information.[44] Estimates greater than 0.9 tend to indicate situations where the covariate distributions are quite similar, so that reweighting may not even be necessary. Estimates between these two extremes may suffer from increased variance (due to large weights) or bias (due to poor overlap in the finite sample). We simulated three scenarios with small to moderate differences in the characteristic distributions.

Among the three selection mechanisms introduced earlier, Scenario 1 represents small imbalance ($\Delta_p \approx 0.05$ and $\mathcal{B}_T \approx 0.96$) due to a small selection effect in four out of the 16 characteristics. Scenario 2 represents moderate imbalance ($\Delta_p \approx 0.31$ and $\mathcal{B}_T \approx 0.76$) due to larger selection effect in four characteristics and smaller effect for the remaining ones. Finally, in Scenario 3, we again have moderate imbalance ($\Delta_p \approx 0.28$ and $\mathcal{B}_T \approx 0.76$), but this selection effect is entirely driven by a single, strong characteristic.

In addition, we use two model summary metrics, R^2 and coefficient of variation (CV), to give some further intuition to these different outcome scenarios, and report these in Table 2. The formulas for calculating these two metrics are shown below:

$$R^2 = \frac{\text{Var}\left(\mu + \gamma Z_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\theta}^T \boldsymbol{X}_i Z_i\right)}{\text{Var}\left(\mu + \gamma Z_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\theta}^T \boldsymbol{X}_i Z_i + \varepsilon_i\right)} = \frac{\frac{4}{25}\gamma^2 + \frac{1}{12}\sum_{i=1}^{p}\beta_j^2 + \frac{17}{300}\sum_{i=1}^{p}\theta_j^2}{\frac{4}{25}\gamma^2 + \frac{1}{12}\sum_{i=1}^{p}\beta_j^2 + \frac{17}{300}\sum_{i=1}^{p}\theta_j^2 + \sigma^2}, \text{CV} = \frac{\sqrt{\text{Var}\left(\gamma + \boldsymbol{\theta}^T \boldsymbol{X}_i\right)}}{E\left(\gamma + \boldsymbol{\theta}^T \boldsymbol{X}_i\right)} = \frac{\sqrt{\frac{1}{12}\sum_j \theta_j^2}}{\gamma + \frac{1}{2}\sum_j \theta_j}. \quad (12)$$

While the CV measure greatly depends on the baseline treatment effect $\gamma = 2$, we find it helpful for improving our intuition. For example, out of the four outcome scenarios, Scenarios A and B have comparable R^2, but Scenario B has smaller treatment effect variability, as represented by a percentage of the mean, since the source of heterogeneity is spread evenly across many variables. Scenarios C and D have more treatment effect heterogeneity than the other scenarios, as the source of heterogeneity is both strong (i.e., has a large regression coefficient) and focused (i.e., only involves one characteristic).

Lastly, the primary measure of performance of the reweighting estimator is mean squared error (MSE). At each repetition, the estimated E(TATE) from different estimators will be compared against the true value as specified per outcome mechanism and Equation (10). After all the repetitions are completed, MSE will be computed as well as its decomposition: squared bias and variance. To evaluate the validity of the nonparametric bootstrap variance estimate, we further calculate the 95% confidence interval within each repetition and summarize the empirical coverage percentage.

3.1.3 | Models being compared

There are in total 18 estimates (three reweighting estimators by two weight calculations by three weight trimmings) per combination of selection and outcome scenario. The primary focus is to identify the optimal, if any, reweighting estimator and weight calculator across various DGMs. The risk–benefit profile of weight trimming comes secondary. The results will be presented first by selection mechanism and then sub-stratifying by outcome mechanism.

To provide a benchmark performance, we consider an additional estimator based on the true propensity score. Specifically, we plug the true propensity based on Equation (4) into Equations (3) and (6) to compute subject weights and use these in the nonparametric estimator in Equation (8).

3.1.4 | Other implementation details

An R package ebal[36] was utilized to calculate the entropy weights that minimizes the objective function in Equation (5), but the constraints are limited to balance the first moment. The simulation code can be provided upon request.

3.2 | Simulation results

We present the MSE, and its decomposition into variance and squared bias, of the proposed estimators for each selection model in Figures 1–3, one for each selection model. Each quadrant of the figure represents one outcome model, and different estimator-by-weight combinations are shown across bars within a quadrant. Because the results for the nonparametric and parametric-trial estimators, and the results for the two trim cutoffs C are similar, we relegate the presentation of the latter of each pair to the Appendix (Figures S1–S3). The total height of each bar indicates the MSE, which is further decomposed into variance and squared bias, as color coded in green and blue, respectively.

We first consider estimation with untrimmed weights (the second, third, sixth and seventh bars of each figure). Across all combinations of selection-by-outcome scenarios, we observe approximately unbiased estimates, indicating the proposed procedure accurately generalizes the treatment effect observed in a RCT to a target population. Entropy-based weighting appears to provide smaller variance when a nonparametric estimator is used, whereas a propensity score-based method has an advantage when a parametric estimator is implemented. It is also evident that the parametric estimators are in general more efficient than the nonparametric one, which is expected given that parametric models were used to generate the simulated data. However, caution should be taken and model diagnostics should be carried out whenever a parametric model is used to alleviate the concern of model mis-specification.[46] Another common observation is that the MSE of the proposed estimators is either on par with or smaller than that of benchmark performance, where the known PS is used in the nonparametric estimator. This finding is reminiscent of efficacy gain using propensity score estimates rather than true propensity scores.[47,48] The effect of trimming is not uniform across all selection-by-outcome scenarios. As indicated in Figure 1, under Selection Model 1, where there is substantial overlap in the characteristic distributions, the performances across all estimator-by-weight calculation combinations are very similar for the trimmed and untrimmed weights. Given a substantial overlap between trial and target populations, understandably there are not many outlying weights and thus the weight trimming has limited effect in this case.

On the other hand, in Selection Model 2 where there is less overlap, Figure 2 highlights some differences across the simulated combinations. First, trimming reduces a considerable amount of variance. However, the net gain in MSE in

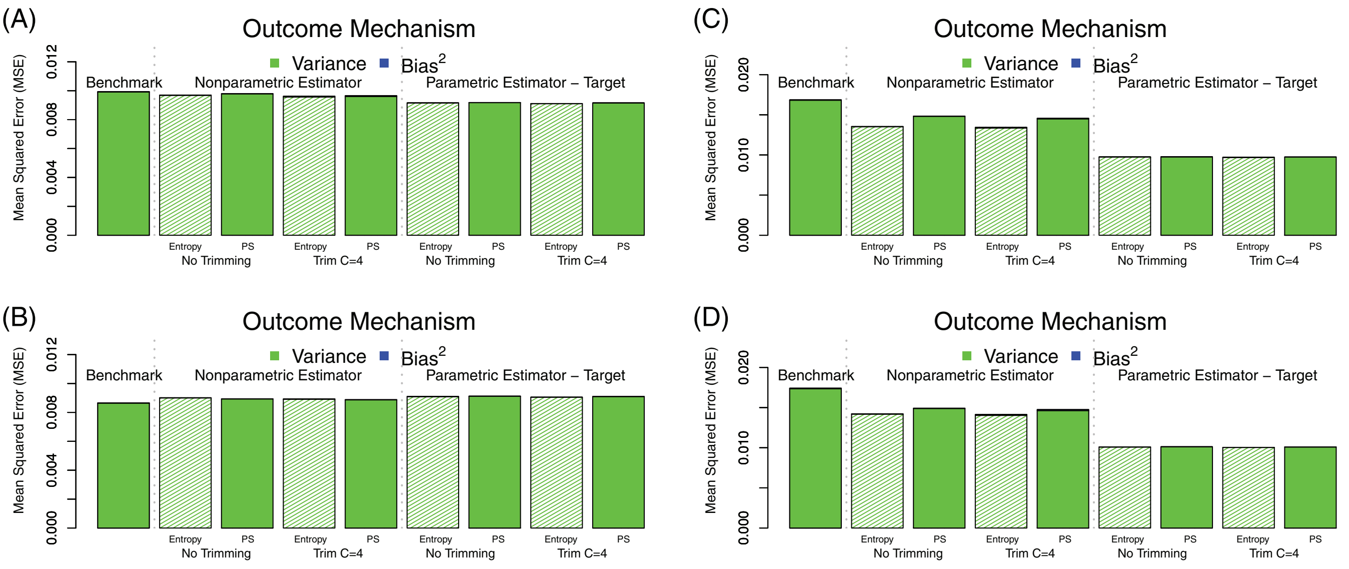

FIGURE 1 Bias and variance of estimators of $E(\text{TATE})$ from selection mechanism 1.
Note: The range of y-axis varies from one outcome mechanism to another. Abbreviations: C, cutoff; PS, propensity score; TATE, target population average treatment effect

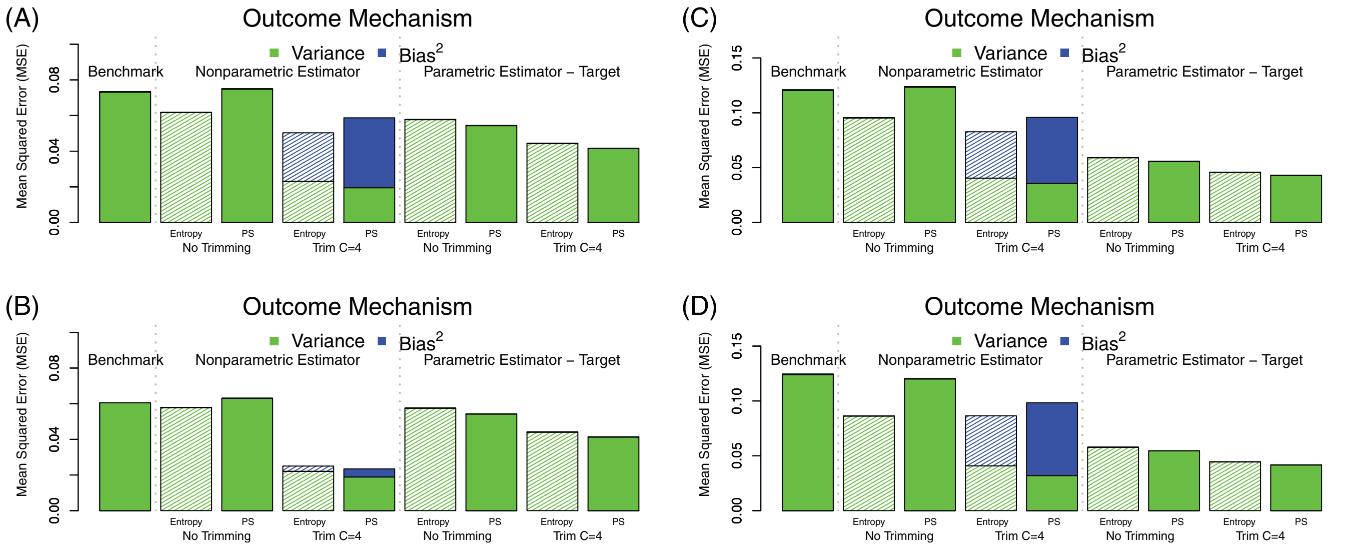

FIGURE 2 Bias and variance of estimators of $E(\text{TATE})$ from Selection mechanism 2.
Note: The range of y-axis varies from one outcome mechanism to another. Abbreviations: C, cutoff; PS, propensity score; TATE, target population average treatment effect

some cases is offset by the large amount of bias it introduces, especially in the presence of treatment effect heterogeneity. It is also noticeable that trimming introduces more bias in the nonparametric estimator than the parametric one. The parametric estimator is arguably less sensitive to weight trimming as a result of an indirect trimming effect on the WLS estimates in Equation (10). On the contrary, the effect of weight trimming on the nonparametric estimator is more directly and explicitly evidenced in Equation (7), making the nonparametric estimator less robust to trimming. Similarly, the weighted RCT characteristics after trimming will no longer be unbiased to the target characteristics, thus making the other parametric estimator that utilizes weighted RCT characteristics ($E(\mathrm{T\hat{A}TE})_{p-\text{trial}}$) behave similarly to the less robust nonparametric estimator, as shown in the Appendix (Figure S2).

Although the degree of characteristic overlap is comparable between Selection Models 2 and 3, the selection effect is singularly driven by the first characteristic in Model 3, resulting in different patterns of MSE. That difference is

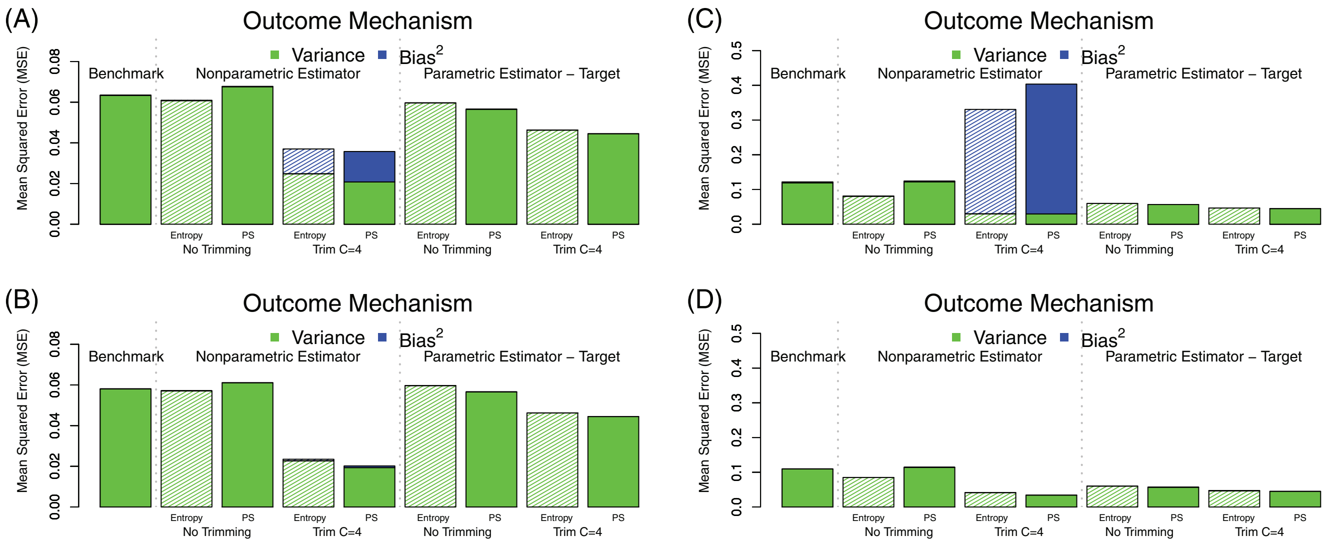

FIGURE 3 Bias and variance of estimators of E(TATE) from Selection mechanism 3.
Note: The range of y-axis varies from one outcome mechanism to another. Abbreviations: C, cutoff; PS, propensity score; TATE, target population average treatment effect

especially noticeable from Outcome Models C and D. Without implementing a trimming strategy, both propensity- and entropy-based methods are able to overcome the selection effect despite the different effect mechanisms, resulting in the very small amounts of bias shown in Figure 3. However, due to a strong, focused selection effect of the first characteristic, subjects with low X_1 values will be heavily weighted, with weight values presumably much larger than the prespecified upper bound C=4. Trimming will not introduce a significant amount of bias if the treatment heterogeneity is mild or moderate as observed in Outcome Models A and B. However, there is a sharp contrast between Selection Model 2 and 3 coupled with Outcome Models C and D. In o Outcome Model C, where the characteristic that drives selection bias coincides with the one that drives treatment heterogeneity, trimming results in nonparametric estimators with considerable amounts of bias that are three to four times greater than the variance trimming reduces. On the other hand, when the characteristics that account for selection bias and treatment heterogeneity are different, trimming becomes a powerful tool to reduce variance without trading off much bias. (Note that the panels in Figure 3 have markedly different scales on the vertical axes). This example illustrates the multitude of factors to consider when generalizing the treatment effect observed in a RCT to a target population, including not only patient characteristic differences and potential treatment heterogeneity, but also the synergistic effect between the two.

These patterns in the effect of weight trimming are similar for the 99th percentile trim, as shown in the Appendix (Figures S1–S3). As expected, the pattern is slightly more pronounced for the percentile trim for the Selection Model 1 simulations, where few of the small individual weights would exceed the fixed $C = 4$ trimming threshold, but 1% of the weights are still trimmed in the percentile trim. The pattern is less pronounced in Selection Model 2 and 3 simulations, where more than 1% of the estimated weights would exceed the $C = 4$ trimming threshold.

Through simulations, we also evaluate bootstrap variance estimation. The empirical SD out of 500 replicates is presented side-by-side with the average SE estimate in Table 3. In most cases, the bootstrap method provides a comparable estimate, though fairly consistently slightly under-estimates the empirical standard deviation. In Outcome Models C and D, the bootstrap method provides a more starkly under-estimated variability particularly for the propensity-based estimates. This subsequently impacts the CI coverage based on normal approximation as shown in Table 4. Conversely, constructing a CI by quantiles is more robust and the coverage rate is close to the nominal rate of 95% in all scenarios (see Table S2).

4 | REAL DATA EXAMPLE

To examine performance outside the realm of simulated data, we apply the proposed estimation procedures to data collected from two clinical trials in patients with mild-to-moderate Alzheimer's disease. Contrary to the original purpose

TABLE 3 Variance estimate of the nonparametric estimator of *E*(TATE)

		Outcome mechanism			
Selection mechanism	**Balancing method**	**Scenario A**	**Scenario B**	**Scenario C**	**Scenario D**
1	PS	0.099/0.101	0.095/0.096	0.122/0.125	0.122/0.126
	Entropy	0.098/0.101	0.095/0.096	0.116/0.125	0.119/0.126
2	PS	0.274/0.228	0.251/0.213	0.352/0.291	0.347/0.294
	Entropy	0.249/0.228	0.241/0.213	0.309/0.291	0.294/0.294
3	PS	0.260/0.237	0.247/0.224	0.350/0.320	0.338/0.307
	Entropy	0.247/0.237	0.239/0.224	0.284/0.320	0.292/0.307

Note: Shown in each cell is SD $(\widehat{\text{tPATE}})/\text{Avg}\ \widehat{\text{SE}}(\widehat{\text{tPATE}})$.

Abbreviations: PS, propensity score; SD, standard deviation; SE, standard error; TATE, target population average treatment effect.

TABLE 4 Empirical coverage of the 95% confidence interval of nonparametric estimator of *E*(TATE)

		Outcome mechanism			
Selection mechanism	**Balancing method**	**Scenario A**	**Scenario B**	**Scenario C**	**Scenario D**
1	PS	0.956/0.956	0.952/0.95	0.974/0.970	0.960/0.960
	Entropy	0.952/0.956	0.956/0.95	0.966/0.970	0.962/0.960
2	PS	0.908/0.916	0.938/0.94	0.902/0.910	0.906/0.912
	Entropy	0.908/0.916	0.902/0.94	0.928/0.910	0.928/0.912
3	PS	0.946/0.944	0.940/0.948	0.892/0.906	0.920/0.926
	Entropy	0.940/0.944	0.932/0.948	0.948/0.906	0.950/0.926

Note: Shown in each cell is the empirical coverage of 95% CI constructed by normal approximation and quantiles.

Abbreviations: CI, confidence interval; PS, propensity score; TATE, target population average treatment effect.

of generalizing a treatment effect observed in a RCT to a real world target population, this real data analysis treats patients from a second clinical trial as the target population. As such, our analysis represents an idealized situation in which the data collection in these two studies is very consistent, the patient- and population-level data are both available and, most importantly, efficacy endpoints are available in both studies. The availability of efficacy endpoints in both studies allows us to showcase additional evidence to support the performance of the proposed estimators outside of the simulations described above.

In two phase 3 clinical trials, EXPEDITION1 and EXPEDITION2 (NCT00905372 and NCT00904683), patients with mild-to-moderate Alzheimer's disease were randomized to receive either solanezumab, a humanized monoclonal antibody that binds to amyloid beta, or placebo. Study results were previously published.[49] Briefly, both studies failed to meet their respective primary endpoints at week 80: change in an 11-item cognitive subscale of the Alzheimer's Disease Assessment Scale (ADAS-Cog11) in EXPEDITION1, and change in the full ADAS-Cog14 in the subset of patients with mild Alzheimer's disease in EXPEDITION2. Due to the nature of the clinical trial, the participants of these two studies did not consent for their data to be shared publicly, and supporting data is not available.

In this real data analysis, we focus on estimating the effect of solanezumab on ADAS-Cog14. We use the full mild-to-moderate Alzheimer's disease patient population from EXPEDITION1 as the RCT population and those from EXPEDITION2 as the target population. While it is unusual to choose the participants of a second trial to be the target population, this choice could be useful for designing follow-up trials. For purposes of illustration, this choice more importantly provides the "true" effect so that we can concretely evaluate our generalization methods. In EXPEDITION1, 1012 patients were randomized in a 1:1 ratio to solanezumab ($N = 506$) and placebo ($N = 506$). Baseline demographics and clinical characteristics were well balanced between treatment groups. Compared to the RCT population, patients in the target population are on average 2 years younger, less frequently have a family history of Alzheimer's disease, have less severe disease as measured by Mini–Mental State Examination (MMSE) and ADAS-Cog14, and are more likely to use an acetylcholinesterase inhibitor (AchEI) as mono-therapy. Gender, years of education and ApoE genotype distributions are very similar. More detailed baseline characteristics can be found in the report of the solanezumab phase 3 trial results by Doody et al.[48] Due to the nature of clinical trial, participants of these two studies did not consent for their data to be shared publicly, so supporting data is not available.

In the propensity and entropy models, all the aforementioned characteristics plus duration of disease diagnosis are included, except for APoE status due to the large amount of missing values. Continuous characteristics are standardized with sample mean 0 and standard deviation 1 across both studies before model fitting. Odds ratio estimates from the logistic model are shown in Figure 4. Based on the propensity score, the Tipton index $\mathcal{B}_T$ and Δ_p are 0.93 and 0.12, respectively, indicating fairly similar populations and substantial overlap in EXPEDITION1 and EXPEDITION2. This is confirmed visually by the histogram in Figure 5, showing the overlap in propensity scores between the two populations.

We apply four generalization estimators, parametric versus non-parametric using propensity- versus entropy-based weights, to the EXPEDITION1 study. In both propensity and entropy models, patients receiving placebo and solanezumab are pooled to build a single analytical model. For the parametric estimator, the baseline characteristics observed in EXPEDITION2 are used. Weight trimming is not considered in this real data implementation given the substantial degree of overlap evidenced by $\mathcal{B}_T$ and Δ_p. No meaningful covariate imbalance was found after weighting.

The raw treatment effect and generalized effect estimates are presented in Figure 6. The raw treatment effects from EXPEDITION1 and EXPEDITION2, as measured by mean difference in ADAS-Cog14 between the solanezumab and placebo groups, are −1.33 and −0.98, respectively. By applying the weighting procedure, the estimated target population average treatment effect ranges from −0.90 to −1.11. All four estimators coherently indicate a smaller treatment effect in the target EXPEDITION2 population as compared to the EXPEDITION1 trial results, consistent with the raw

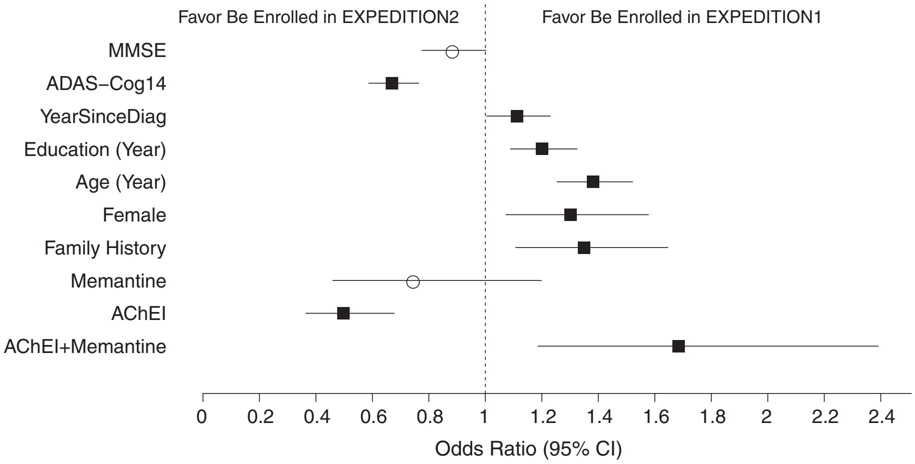

FIGURE 4 Forest plot of logistic model. Abbreviations: AChEI, acetylcholinesterase inhibitor; ADAS-Cog14, 14-item Alzheimer's Disease Assessment Scale-Cognitive subscale; CI, confidence interval; MMSE, Mini-Mental State Examination

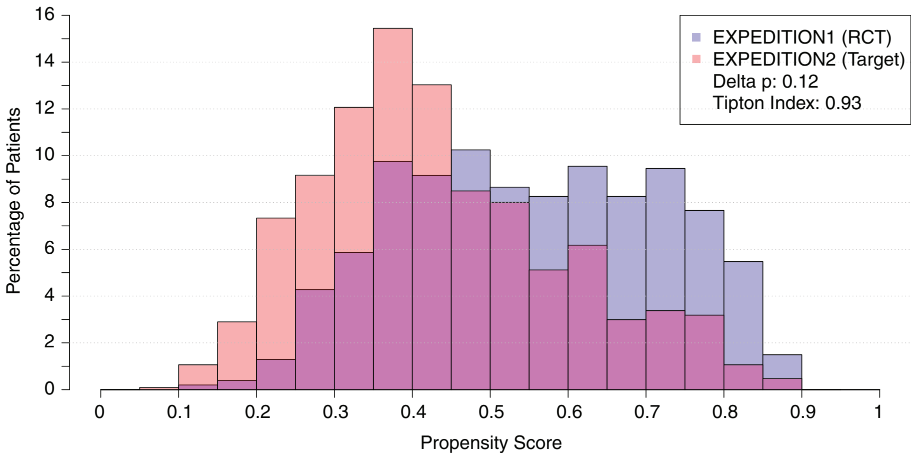

FIGURE 5 Histograms of propensity score overlap between EXPEDITION1 and EXPEDITION2. Abbreviation: RCT, randomized controlled trial

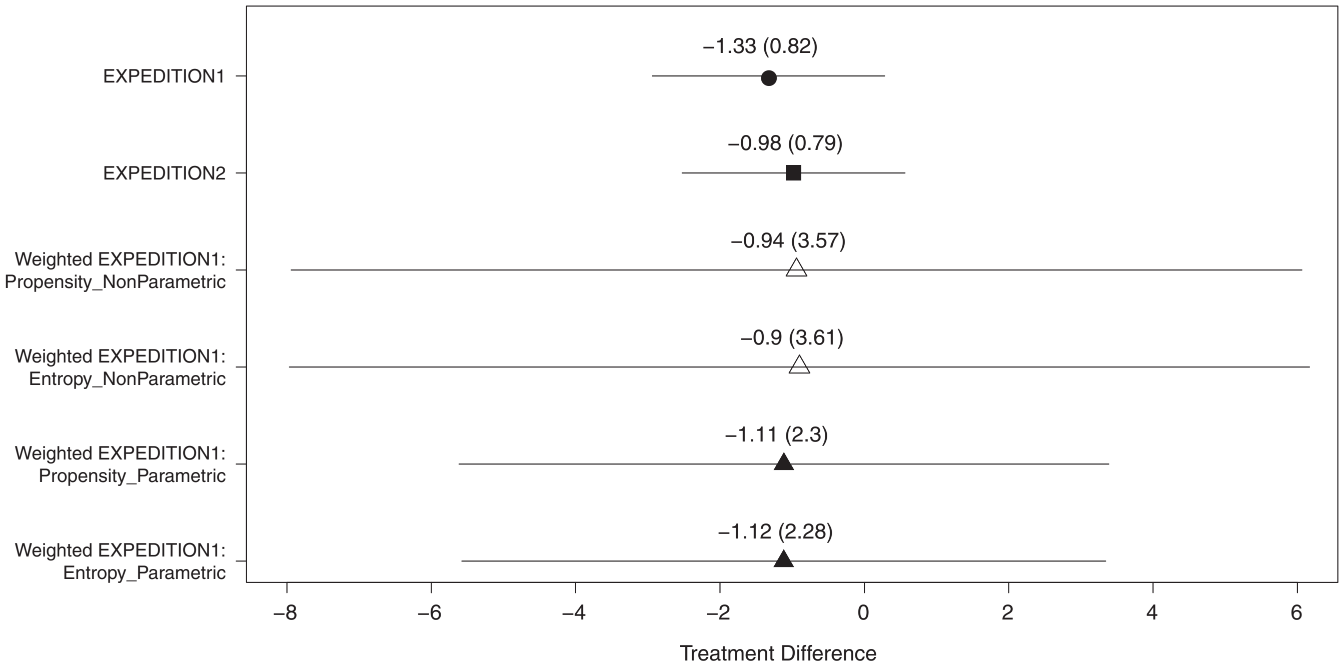

FIGURE 6 Treatment effect estimate from EXPEDITION1/2 and weighted treatment effect estimate from EXPEDITION1. Shown in the figure are point estimate (standard error) and 95% confidence interval

treatment effect observed in the target population. Among them, the non-parametric estimator shrinks the effect more aggressively. Similar estimates are calculated from propensity and entropy weight approaches. Notably, the uncertainty of the generalization estimator, as summarized by the reported confidence interval, is quite large as compared to the uncertainty of the EXPEDITION1 trial results. This result is expected, as the generalization estimator incorporates the uncertainty from both sampling and estimation, and the weighting procedure reduces the effective sample size of the RCT in proportion to the imperfect overlap of propensity score distributions (as shown in Figure 5).

In this real data analysis, we demonstrate the utility of the proposed weighting procedure beyond simulated data. The logistic model estimate (Figure 4) suggests selection bias (bias due to cross-study characteristic imbalance) is not driven by a few, strong covariates, but spread across multiple mild-to-moderate covariates. In addition, the parametric outcome model estimates (not shown) indicate small, non-significant treatment heterogeneity, except for memantine monotherapy and years since diagnosis. These features share some commonality with simulation Selection Model 1 and Outcome Model C or D. Consistent with what we learn from simulation, there is no discernible difference between propensity and entropy methods, albeit the entropy method is slightly more efficient when the nonparametric estimator is employed. Meanwhile, the parametric method remains more efficient regardless of balancing method used.

5 | DISCUSSION

We study trial generalization in the context of extending the treatment effect observed in a RCT to a target population. The extension is done by weighting the subjects observed in a RCT so that the weighted trial population is more similar to the target. The estimation procedure performs well in both simulations and real data analyses. Two approaches to calculate patient weight, a model-based propensity method and a direct weight calculation, result in comparable performance in our settings. Given that subject-level data are not always available, this finding provides reassuring preliminary evidence that population-level summary data can be just as accurate and efficient as subject-level data. Though not a common practice, availability of subject-level data allows the model-based propensity model to take full advantage of the joint distribution by incorporating high-order terms in the model. The entropy algorithm can, in theory, also balance higher-order terms by including additional constraints. However, higher-order summary data are less frequently reported when only summary data are available.

Weight calculation is just one of many components to consider in the estimation procedure. Other considerations include, among others, the type of model used to describe differences in patient characteristics, characteristic selection for the propensity/entropy model, and whether and how to trim weights. We used a logistic model for its popularity and simplicity, but more advanced machine learning algorithms have been proposed and demonstrated to be advantageous.[50] Future work can explore novel balancing methods and investigate the benefit of balancing not only the first moment but higher-order moments and interactions. The limitations of an entropy model approach should also be studied to understand when population-level data may not be adequate. Also, for simplicity, we did not conduct characteristic selection in weight calculation. This can be less of a concern to achieve a good balance, but it could be at a cost of higher variability. Finally, weight trimming by itself warrants further investigation. We consider a fixed upper bound and a percentile-based trimming strategy in the simulations, and the result varies from one to another. Moreover, our simulations also suggest the success of trimming depends on the characteristics that contribute either to selection bias, treatment heterogeneity or both. A flexible trimming strategy should be developed and tested so that the degree of trimming can be adapted on a case-by-case basis. Development of diagnostic tools to guide weight adjustment decisions in a principled manner would further help practitioners make sound choices related to weight adjustments.

In this study, we also conclude the estimated propensity is more efficient than the known propensity. As counter-intuitive as it seems that knowing the truth is not as efficient as estimating it, this finding mirrors earlier results by Rosenbaum[47] and Rubin and Thomas.[48] Although this finding gives some reassurance, one caveat to keep in mind is that any estimation carries the risk of violating certain underlying assumptions (e.g., the unmeasured confounder assumption).[14,29] Our simulations did not assess the effects of such violations.

Despite the encouraging outcomes in simulations and real data analysis, there are some methodological limitations that have not been fully addressed in the current study. For instance, we established the effectiveness of the weighting method but did not compare it to other competing methods, such as targeted maximum likelihood estimation[51] or Bayesian additive regression trees.[52] Moreover, while the target population used in our real data analysis is clearly defined, the definition and specificity of a target population may not be well described in real practice settings. Insurance and registry databases can be used to define the profile of a target population, but these databases are also subject to their own selection biases. Finally, the inclusion/exclusion criteria imposed in clinical trials inevitably create a barrier from the real world population. For instance, Malatestinic et al.[53] reported 28.7% of psoriasis patients in the U.S. Department of Defense healthcare database were not eligible to enroll in clinical trials due to the inclusion/exclusion criteria commonly implemented in psoriasis studies. Attempting to extrapolate trial results beyond the patient population that is representative of the target population could introduce model bias and inflate variability. From practicality point of view, data collected in a routine practice could fundamentally differ from that in trail setting due to mandated participation and drug adherence. More importantly, critical confounding factors, such as clinical or disease-specific characteristics, are less routinely collected in real world databases, like insurance claims.

In summary, we have found simple generalization estimators to have great practical potential. In exploring several options for constructing such estimators, we find that propensity-based weights are intuitive and synergistic with parametric estimators (particularly when the parametric models are correctly specified), and entropy-based weights are more adaptable to limited data availability and perform well when utilized in nonparametric estimators. In both cases, we should avoid weight trimming when treatment effect heterogeneity and selection are strongly influenced by the same set of covariates. Future development of diagnostic tools may identify such situations when weight trimming safely provides efficiency gains.

CONFLICT OF INTEREST

Chen-Yen Lin, Douglas Faries and Joseph Johnston are full time employee of the funding source of this research, Eli Lilly and Company. Eloise Kaizar is a scientific consultant of for Eli Lilly and Company.

DATA AVAILABILITY STATEMENT

Due to the nature of clinical trial, participants of these two studies did not consent for their data to be shared publicly, supporting data is not available.

ORCID

Chen-Yen Lin https://orcid.org/0000-0003-0662-4079

REFERENCES

1. Frangakis C. The calibration of treatment effect from clinical trials to target populations. *Clin Trials*. 2009;6(2):136-140.
2. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol*. 2010;172(1):107-115.
3. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference*. 2013;1(1):107-134.
4. Tipton E. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J Educ Behav Stat*. 2013;38(3):239-266.
5. Hartman E, Grieve R, Ramsahai R, Sekhon J. From sample average treatment effect to population average treatment effect on the treated combining experimental with observational studies to estimate population treatment effects. *J R Stat Soc Ser A Stat Soc*. 2015; 178:757-778.
6. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*. 2017;186(8):1010-1014.
7. Philippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018;38(2):200-211.
8. Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernan MA. Generalizing causal inference from individual in randomized trials to all trial-eligible individuals. *Biometrics*. 2018;10:1-12.
9. Buchanan AL, Hudgens MG, Cole SR, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J R Stat Soc Ser A Stat Soc*. 2018;181(4):1193-1209.
10. Dahabreh IJ, Robertson SE, Stuart EA, Hernan MA. Transporting inferences from a randomized trial to a new target population. *Stats Med*. 2020;39(14):1999-2014.
11. Garner S, Thwaites R. Workpackage 1 (Innovative Medicines Initiative [IMI], Get Real) [Online]. https://www.imi-getreal.eu/About-GetReal/Workpackage-1. Cited March 25, 2020.
12. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials. Race-, sex-, and age-based disparities. *JAMA*. 2004;291(22): 2720-2726.
13. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Design for Generalized Casual Inference*. Boston, MA: Cengage Learning; 2002.
14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
15. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;68(6):656-664.
16. US General Accountability Office. *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research*. Washington, DC: US Government Accountability Office; 1992.
17. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Tech Assess Health Care*. 1993;9(3):440-449.
18. Kaizar EE. Estimating treatment effect via simple cross design synthesis. *Stat Med*. 2011;30:2986-3009.
19. Varadhan R, Henderson N, Weiss C. Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogeneous: part I methodology. *Commun Stat*. 2016;2:112-126.
20. Henderson N, Varadhan R, Weiss C. Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogenous: part II. Application and external validation. *Commun Stat*. 2017;3:7-20.
21. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34:3661-3679.
22. DuMouchel W, Duncan G. Using sample survey weights in multiple regression analyses of stratified samples. *J Am Stat Assoc*. 1983;78: 535-543.
23. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846-866.
24. Hirshberg DA, Zubizarreta JR. On two approaches to weighting in casual inference. *Epidemiology*. 2017;28(6):812-816.
25. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating casual effect in observational studies. *Psychol Methods*. 2004;9(4):403-425.
26. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29:337-346.
27. Imai K, Ratkovic M. Covariate balancing propensity score. *J R Stat Soc Series B Stat Methodol*. 2014;76:243-263.
28. Deville JC, Sarndal CE. Calibration estimators in survey sampling. *J Am Stat Assoc*. 1992;87:376-382.
29. Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol*. 2010;172(7):843-854.
30. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *Public Lib Sci One*. 2011;6(3):e18174.
31. Potter F. A study of procedures to identify and trim extreme sampling weights. *JSM Proceedings, Survey Research Methods Section*. Alexandria, VA: American Statistical Association; 1990:225-230.
32. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:945-960.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2018.
34. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit Anal*. 2011;20(1):25-46.
35. Kullback S. *Information Theory and Statistics*. New York, NY: Wiley; 1959.

36. Hainmueller J. ebal: Entropy reweighting to create balanced samples. R-Project Web site. 2019. https://cran.r-project.org/web/packages/ebal/ebal.pdf. Accessed September 13, 2019.
37. NAEP. NAEP Technical documentation trimming of student weights [Online]. 2008. https://nces.ed.gov/nationsreportcard/tdw/weighting/2002_2003/weighting_2003_studtrim.aspx
38. Chen TC, Park JD, Clark J, et al. National health and nutrition examination survey: estimation procedures, 2011–2014. *Vital Health Stat*. 2018;2(177):1-26.
39. Centers for Disease Control and Prevention. *National Immunization Survey-Child: A User's Guide for the 2017 Public-Use Data File*. Chicago, IL: NORC at the University of Chicago; 2018.
40. Pfeffermann D. The role of sampling weights when modeling survey data. *Int Stat Rev*. 1993;61(2):317-337.
41. Muller UU. Weighted least squares estimators in possibly misspecified nonlinear regression. *Metrika*. 2007;39(2007):66.
42. Chen Z, Kaizar E. On variance estimation for generalizing from a trial to a target population. arXiv. 2017. https://arxiv.org/abs/1704.07789. Accessed September 13, 2019.
43. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity score to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc*. 2011;174(2):369-386.
44. Tipton E. How generalizable is your experiment? An index for comparing experimental samples and populations. *J Educ Behav Stat*. 2014;39(6):478-501.
45. Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhya Ind J Stat*. 1946;7:401-406.
46. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523-539.
47. Rosenbaum P. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387-394.
48. Rubin D, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996;52:249-264.
49. Doody RS, Thomas RG, Farlow M, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med*. 2014;370(4):311-321.
50. 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.
51. Schuler M, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65-73.
52. Hill J. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217-240.
53. Malatestinic W, Nordstrom B, Wu J, et al. Characteristics and medication use of psoriasis patients who may or may not qualify for randomized controlled trials. *J Manag Care Special Pharm*. 2017;23(3):370-381.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lin C-Y, Kaizar E, Faries D, Johnston J. A comparison of reweighting estimators of average treatment effects in real world populations. *Pharmaceutical Statistics*. 2021;20:765–782. https://doi.org/10.1002/pst.2106