Extending inferences from a randomized trial to a new target population

Issa J. Dahabreh MD ScD[1-4] | Sarah E. Robertson MS[1,2] |
Jon A. Steingrimsson PhD[5] | Elizabeth A. Stuart PhD[6] |
Miguel A. Hernán MD DrPH[4,7,8]

[1]Center for Evidence Synthesis in Health, Brown University, Providence, RI

[2]Department of Health Services, Policy & Practice, Brown University, Providence, RI

[3]Department of Epidemiology, Brown University, Providence, RI

[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

[5]Department of Biostatistics, School of Public Health, Brown University, Providence, RI

[6]Departments of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[8]Harvard-MIT Division of Health Sciences and Technology, Boston, MA

Correspondence
Issa Dahabreh MD ScD, Box G-S121-8; Brown University, Providence, RI 02912.
Email: issa_dahabreh@brown.edu

When treatment effect modifiers influence the decision to participate in a randomized trial, the average treatment effect in the population represented by the randomized individuals will differ from the effect in other populations. In this tutorial, we consider methods for extending causal inferences about time-fixed treatments from a trial to a new target population of non-participants, using data from a completed randomized trial and baseline covariate data from a sample from the target population. We examine methods based on modeling the expectation of the outcome, the probability of participation, or both (doubly robust). We compare the methods in a simulation study and show how they can be implemented in software. We apply the methods to a randomized trial nested within a cohort of trial-eligible patients to compare coronary artery surgery plus medical therapy versus medical therapy alone for patients with chronic coronary artery disease. We conclude by discussing issues that arise when using the methods in applied analyses.

KEYWORDS
transportability, generalizability, randomized trials, observational analyses, double robustness

Abbreviations: DR, doubly robust; IO, inverse odds; OM, outcome model-based.

1 | BACKGROUND

Participants in randomized trials often differ from individuals seen in clinical practice with respect to the distribution of covariates that modify the treatment effect. Because of such differences, the average treatment effect in the population of trial participants differs from the average treatment effect in other populations [1]. This lack of generalizability or transportability of trial results has been documented by numerous surveys comparing the distribution of baseline covariates between trial participants and samples of patients seen in clinical practice (e.g., [2, 3, 4]).

The task of assessing whether trial findings are relevant to a new target population is often handled via informal subjective assessments [5]. Clinicians and policy makers, however, cannot easily use such assessments to draw inferences about the target population they are interested in (e.g., the population of patients in their practice or the population to which the policy will be applied) and they cannot evaluate the performance of informal assessments for guiding decisions.

This tutorial builds on a growing literature (e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]) to show that, instead of relying exclusively on informal assessments, we can use formal statistical methods to extend causal inferences from trial participants to a new target population. We focus on the common setting in which individual-level data on time-fixed treatments, outcomes, and baseline covariates are available from a randomized trial, but only individual-level data on baseline covariates are available from the target population of non-participants. We examine methods for extending inferences using models for the expectation of the outcome, the probability of trial participation, or both (doubly robust). We assess the finite-sample performance of the methods in simulation studies and provide example code to implement the methods in R. We apply the methods to a randomized trial nested within a cohort of eligible patients with coronary artery disease that compared coronary artery surgery plus medical therapy versus medical therapy alone. Last, we discuss issues that arise when using the methods in practice.

2 | EXTENDING TRIAL FINDINGS TO A TARGET POPULATION

2.1 | Basic concepts and terminology

Suppose that we have access to data from a completed trial that compared two or more treatments. We want to evaluate the comparative effectiveness of the treatments studied in the trial in some well-defined *target population of non-participants*. Usually, the distribution of effect modifiers differs between the population represented by trial participants and the target population. Such differences are due to the trial participation process [17]: First, the investigators specify eligibility criteria, which define the *trial-eligible population*. Next, they invite eligible individuals to participate in the trial and the recruitment activities define the *invited population*. Last, the invited individuals who chose to participate in the trial define the *participant population*.

We use the term *generalizability* when the target population coincides or is a subset of the trial-eligible population and *transportability* when the target population includes at least some individuals who are not trial-eligible (and who, by definition, cannot be trial participants) [17, 18] (others have proposed different definitions [14]). In generalizability analyses, the target population often has a different distribution of effect modifiers compared with the participant population [16], even though both populations meet the trial eligibility criteria. In transportability analyses, such differences are more likely.

We collectively refer to generalizability and transportability as *extending inferences from a trial to a target population*. The methods described in this paper apply to analyses extending inferences (i.e., both generalizability and transportability) from a trial to a target population of non-participants. Henceforth, for brevity, we mostly use the term "target

population" without the qualifier "of non-participants;" it should be understood, however, that our results are limited to that population.

2.2 | Study designs

The most natural study design for extending inferences from a randomized trial to the target population is a *nested trial design*, in which the randomized trial is embedded within a simple random sample of individuals from a broader population, including individuals who refuse to participate in the trial and those who are not invited [19]. For example, in comprehensive cohort studies [20], a trial is nested within a cohort of trial-eligible individuals. In this design, investigators collect baseline covariate information from all cohort members, but need only collect treatment and outcome information from randomized individuals. This setup also occurs in trials embedded in health-care systems where baseline covariate data are routinely collected from all members of the system, but only a subset of the members participate in the trial [21].

An alternative for extending inferences to the target population is a *non-nested trial design*, in which the randomized trial is combined with a separately obtained sample from the target population [19]. Here, investigators create a *composite dataset* by appending the data from a completed trial to a separately obtained sample from the target population (e.g., [14]); the composite dataset facilitates modeling and estimation, as we discuss below. We assume that the sampled non-participants are a simple random sample from the target population. More specifically, we assume that the trial sample and the sample of non-participants are obtained separately; and that all trial participants, but only a subset of non-participants, are sampled (the latter with unknown sampling probability); see Section B of the Appendix for details about the underlying sampling scheme. As in nested trial designs, baseline covariate information is available both from randomized and non-randomized individuals, but treatment and outcome information is only available from randomized individuals. This setup occurs in drug development and regulatory settings, because treatment and outcome data are available only from a small number of trial participants prior to drug approval, but baseline covariate data can be collected from large samples of individuals who would be eligible for participation in the trials and can be identified in routinely collected data (for treatments that are commercially available, trial-eligible new-users [22] of the treatments compared in the trials may be a reasonable choice of non-participants). The setup also occurs in policy research, when randomized trials are conducted in selected samples but target population data are available from administrative databases or surveys.

The methods we describe in this tutorial can be used both in nested and non-nested trial designs to estimate treatment effects in the target population, using baseline covariate, treatment, and outcome data from the trial and only baseline covariate data from the target population. Treatment and outcome data from the target population, if available, can be used to evaluate assumptions [11, 16], but they are not necessary for the types of analyses described in this tutorial. This is an attractive feature if treatment and outcome data from non-participants are unreliable (e.g., due to confounding of the treatment-outcome association or gross measurement error), costly to obtain, or impossible to collect.

3 | DATA AND CAUSAL QUANTITIES OF INTEREST

The data generated from the designs described in the previous section consist of independent observations on baseline covariates X; assigned treatment A; outcome Y; and a trial participation indicator S that takes the value 1 for trial participants or 0 for non-participants. In nested trial designs, the data consist of realizations of independent and

identically distributed random tuples $(X_i, S_i, S_i \times A_i, S_i \times Y_i)$, $i = 1, \ldots, n$, where n is the total number of observations and we define $n_{\text{RCT}} = \sum_{i=1}^{n} I(S_i = 1)$ and $n_{\text{obs}} = \sum_{i=1}^{n} I(S_i = 0)$, where $I(\cdot)$ denotes the indicator function. In non-nested trial designs, the data available from the randomized trial are realizations of $(X_i, S_i = 1, A_i, Y_i)$, $i = 1, \ldots, n_{\text{RCT}}$; the data from the sample of the target population are realizations of $(X_i, S_i = 0)$, $i = 1, \ldots, \widetilde{n}_{\text{obs}}$ and we define the total number of (sampled) observations that contribute data to the analysis as $\widetilde{n} = n_{\text{RCT}} + \widetilde{n}_{\text{obs}}$. Note that the number of trial participants, n_{RCT}, is the same as in nested trial designs, because in non-nested trial designs all trial participants are sampled; but $\widetilde{n}_{\text{obs}} \neq n_{\text{obs}}$ and $\widetilde{n} \neq n$, because only a subset of the non-participants are sampled (with unknown sampling probability).

In both designs, the data exhibit a special missingness pattern: for trial participants we have data on $(S = 1, X, A, Y)$, but for non-participants we only have data on $(S = 0, X)$. In the main text of the tutorial, we focus on nested trial designs because we think that they will be increasingly used in the future and because our motivating application had such a design; in the Appendix we discuss the sampling properties and give additional results for non-nested trial designs. Table 1 shows the observed data structure for the special case of binary treatment in a nested trial design.

TABLE 1 Data structure for binary treatment, A, observed outcome, Y, and baseline covariates, X, for trial participants ($S = 1$) and non-participants ($S = 0$) in a nested trial design. The observed data include ($S = 1, X, A, Y$) information from n_{RCT} trial participants, of whom n_0 are assigned to $A = 0$ and n_1 assigned to $A = 1$; and ($S = 0, X$) information from n_{obs} non-participants; n is the total number of observations. Dashes denote missing data.

Unit	S	A	Y	X
1	1	0	y_1	x_1
$\vdots$	1	$\vdots$	$\vdots$	$\vdots$
n_0	1	0	y_{n_0}	x_{n_0}
$1 + n_0$	1	1	y_{1+n_0}	x_{1+n_0}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_1 + n_0 = n_{\text{RCT}}$	1	1	$y_{n_1+n_0}$	$x_{n_1+n_0}$
$1 + n_{\text{RCT}}$	0	–	–	$x_{n_1+n_0+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_{\text{obs}} + n_{\text{RCT}} = n$	0	–	–	x_n

We use the random variables Y^a to denote the potential (counterfactual) outcomes under intervention to set treatment to a, possibly contrary to fact [23, 24]. We only consider treatments with a finite number of possible levels in this tutorial; extensions to continuous treatments are possible but not pursued here. For any two treatment values, a, a', we are interested in the average treatment effect in the target population, $E[Y^a - Y^{a'} | S = 0]$. Because the average treatment effect in randomized trials without loss to follow-up, $E[Y^a - Y^{a'} | S = 1]$ is identifiable under standard assumptions [25], it is interesting to compare $E[Y^a - Y^{a'} | S = 1]$ against $E[Y^a - Y^{a'} | S = 0]$. Though often assumed equal, these two effects are different when the effect of treatment varies over baseline covariates that are differentially distributed between trial participants and non-participants (i.e., when the distribution of effect modifiers on the additive scale is different between trial participants and non-participants) [26, 27].

When using nested trial designs, another causal effect, the average effect in the population represented by the cohort in which the trial is nested, $E[Y^a - Y^{a'}]$, may also be of interest, and we have discussed its identification and esti-

mation in previous work [16]. Even if the primary target of inference for nested trial designs is $E[Y^a - Y^{a'}]$, investigators are typically also interested in estimating the effect among non-participants, $E[Y^a - Y^{a'}|S = 0]$. Comparing the effect among non-participants against the effect among trial participants is often of substantial scientific and policy interest. Importantly, when using non-nested trial designs, $E[Y^a - Y^{a'}]$ cannot be identified unless additional information about the sampling of individuals from the target population is available (for more details, see the Appendix of this tutorial, and references [19] and [28]). In the absence of such information, $E[Y^a - Y^{a'}]$ is not identifiable in non-nested trial designs because the data are a "mixture" of trial participants and non-participants, with arbitrary mixing proportions (determined by the size of the trial and the sample from the target population). Thus, in the remainder of this tutorial we focus on the identification and estimation of the components $E[Y^a|S = 0]$ of the average treatment effect in the target population.

4 | IDENTIFIABILITY CONDITIONS & IDENTIFICATION

4.1 | Identifiability conditions

We now discuss sufficient conditions for identifying the mean of the potential outcomes under each treatment a in the target population, $E[Y^a|S = 0]$.

(1) Consistency of potential outcomes: The observed outcome for the ith individual who received treatment a equals that individual's potential outcome under the same treatment, that is, for every i, if $A_i = a$, then $Y_i = Y_i^a$. Implicit in this notation is the assumption that trial participation does not affect the outcomes except through treatment assignment (e.g., there are no Hawthorne effects) [29].

(2) Conditional exchangeability in mean over A in the randomized population: Among trial participants, the potential outcome mean under treatment a is independent of treatment, conditional on baseline covariates: $E[Y^a|X = x, S = 1, A = a] = E[Y^a|X = x, S = 1]$ for every x with positive joint density $f_{X,S}(x, S = 1)$. We expect conditional exchangeability in the trial to hold, regardless of whether the randomization was unconditional or conditional on covariates. The mean exchangeability assumption is weaker than the assumption of exchangeability in distribution over A, that is, $Y^a \perp\!\!\!\perp A|X, S = 1$.

(3) Positivity of treatment assignment: In the trial, the probability of being assigned to each of the treatments being compared, conditional on the covariates needed for exchangeability over A in the randomized population, is positive: $Pr[A = a|X = x, S = 1] > 0$ for every a and every x with positive joint density $f_{X,S}(x, S = 1) > 0$.

(4) Conditional exchangeability in mean over S: The potential outcome mean is independent of trial participation, conditional on baseline covariates: $E[Y^a|X = x, S = 1] = E[Y^a|X = x, S = 0]$ for every x with positive joint density $f_{X,S}(x, S = 0)$. Again, this assumption is weaker than the assumption of exchangeability in distribution over S, that is, $Y^a \perp\!\!\!\perp S|X$.

(5) Positivity of trial participation: The probability of participating in the trial, conditional on the covariates needed to ensure exchangeability over S, is positive, $Pr[S = 1|X = x] > 0$, for every x with positive joint density $f_{X,S}(x, S = 0) > 0$. Informally, this condition ensures that, for the covariates that are necessary to achieve exchangeability over S, all covariate patterns in the target population will be represented in the trial as the trial sample size gets arbitrarily large. Note that the probability of trial participation does not need to be bounded away from 1, since it is not possible to have $f_{X,S}(x, S = 0) > 0$ and $Pr[S = 1|X = x] = 1$ for any x.

Consistency, exchangeability over A in the randomized population, and positivity of treatment assignment are expected to hold for well-defined interventions assessed in (marginally or conditionally) randomized trials. In contrast, positivity of trial participation can be challenging to verify in the data [30] and exchangeability over S is a strong and

untestable assumption. Thus, both positivity of trial participation and exchangeability over S need to be examined on the basis of substantive knowledge when extending inferences from a trial to a new target population.

Now that we have stated the identifiability conditions, we can revisit the relationship of the target population and the trial-eligible population. As we noted in Section 2, in transportability analyses the target population includes individuals who are not trial-eligible. Conditions (4) and (5) suggest that such transportability analyses are possible only to the extent that some of the variables defining the trial eligibility criteria are not needed to ensure exchangebility over S (e.g., maybe the trial restricted eligibility to particular geographic locations but the investigators believe that conditional on other variables location itself is not associated with the outcome). If all variables used to define the trial eligibility criteria are needed for exchangeability over S then the target population has to be limited to *trial-eligible* non-participants.

4.2 | Identification

Under the above assumptions, as shown in the Appendix, for each treatment a, the potential outcome mean in the target population can be identified using the observed data,

$$\mathsf{E}[Y^a|S = 0] = \mathsf{E}\left[\mathsf{E}[Y|X, S = 1, A = a]\big|S = 0\right].$$

The average treatment effect comparing treatments a and a' in the target population can be identified as

$$\mathsf{E}[Y^a - Y^{a'}|S = 0] = \mathsf{E}\left[\mathsf{E}[Y|X, S = 1, A = a]|S = 0\right] - \mathsf{E}\left[\mathsf{E}[Y|X, S = 1, A = a']\big|S = 0\right].$$

In the Appendix we show that the average treatment effect in the target population is identifiable under an assumption of conditional exchangeability in measure over S, which is weaker than conditional exchangeability in mean over S. Under the weaker assumption, we can identify average treatment effects but we cannot identify potential outcome means. Because these means are of inherent scientific and policy interest, in the next section we focus on methods for estimating the functionals $\mu(a) \equiv \mathsf{E}\left[\mathsf{E}[Y|X, S = 1, A = a]\big|S = 0\right]$.

5 | ESTIMATION

We consider three approaches for estimating $\mu(a)$: (1) *outcome modeling*; (2) *probability of trial participation modeling* followed by weighting; and (3) *doubly robust* approaches that combine outcome and trial participation modeling. Under the consistency assumption, these methods can be thought of as solutions to a missing data problem for the outcome, where the missing data indicator is $I(S = 1, A = a)$. In the main text of the tutorial, we focus on developing intuition about the methods; in the Appendix, we provide some theoretical results. Throughout, we assume that the working models are parametric, as is usually the case in applied work (we consider the use of more flexible methods, such as machine learning, in the Discussion).

5.1 | Outcome modeling

The first approach extends inferences from trial participants to the population of non-participants by applying an outcome regression model fit among the former to a sample of the latter [31]. In essence, we use data from trial

participants to estimate models for the expectation of the outcome and then average the model predictions with respect to the covariate distribution of the non-participants. The outcome model-based estimator (OM) of $\mu(a)$ is

$$\widehat{\mu}_{\text{OM}}(a) = \left\{ \sum_{i=1}^{n}(1 - S_i) \right\}^{-1} \sum_{i=1}^{n}(1 - S_i)g_a(X_i; \widehat{\theta}), \tag{1}$$

where $g_a(X; \widehat{\theta})$ is is an estimator for $E[Y|X, S = 1, A = a]$ based on a parametric model with finite-dimensional parameter θ. We usually estimate separate outcome models in each treatment group of the trial, if the data allow, to account for all possible treatment $\times$ covariate interactions. When the model is correctly specified the estimator is consistent for $\mu(a)$.

5.2 | Trial participation modeling and weighting

The second approach extends inferences from trial participants to the population of non-participants using a model for the probability of trial participation [6, 14, 16]. In essence, we are treating the randomized trial participants as a sample from the target population with sampling probabilities that depend on baseline covariates and need to be estimated, an idea that connects this approach to survey sampling [32]. Specifically, we estimate $\mu(a)$ using the IO weighting estimator

$$\widehat{\mu}_{\text{IOW1}}(a) = \left\{ \sum_{i=1}^{n}(1 - S_i) \right\}^{-1} \sum_{i=1}^{n} \widehat{w}_a(X_i, S_i, A_i)Y_i, \tag{2}$$

with the estimated weights $\widehat{w}_a(X_i, S_i, A_i)$ defined as

$$\widehat{w}_a(X_i, S_i, A_i) = \frac{1 - p(X_i; \widehat{\beta})}{p(X_i; \widehat{\beta})e_a(X_i; \widehat{\gamma})} \times I(S_i = 1, A_i = a).$$

Here, $p(X; \widehat{\beta})$ is an estimator for the probability of participation in the trial, $\Pr[S = 1|X]$, based on a parametric model with finite-dimensional parameter β, and $e_a(X; \widehat{\gamma})$ is an estimator for the probability of being assigned to treatment a among trial participants, $\Pr[A = a|X, S = 1]$, based on a parametric model with finite-dimensional parameter γ. The probability of participation in the trial is, in general, unknown and has to be estimated (e.g., by fitting a logistic regression model). In contrast, the probability of treatment in the randomized trial is known (determined by the investigators) and the true values can be used instead of estimated values (estimating the probability may, however, lead to smaller standard errors). We refer to the estimator in (2) as an inverse odds (IO) weighting estimator because $\dfrac{1 - p(X_i; \widehat{\beta})}{p(X_i; \widehat{\beta})}$ is the inverse of the estimated odds of trial participation conditional on baseline covariates.

An alternative IO weighting estimator normalizes the weights to sum to the number of non-randomized individuals,

$$\widehat{\mu}_{\text{IOW2}}(a) = \left\{ \sum_{i=1}^{n} \widehat{w}_a(X_i, S_i, A_i) \right\}^{-1} \sum_{i=1}^{n} \widehat{w}_a(X_i, S_i, A_i)Y_i. \tag{3}$$

The above normalized IO estimator can be obtained using a (saturated) weighted least squares regression of the outcome on treatment, using weights equal to $\widehat{w}_a(X_i, S_i, A_i)$; for example, for binary treatment $\widehat{\mu}_{\text{IOW2}}(a = 0)$ is equal to the intercept of the regression and $\widehat{\mu}_{\text{IOW2}}(a = 1)$ is equal to the sum of the intercept and the coefficient of treatment. Because of the normalization by the sum of the weights, this estimator produces estimates that are always within the support of the outcome random variable (e.g., if Y is binary the estimates are between 0 and 1), resulting in better finite-sample performance when the weights are highly variable. Estimators like this one are sometimes referred to as

"ratio estimators" and have a long history in survey research [33].

When the model for the probability of participation is correctly specified both IO weighting estimators are consistent for $\mu(a)$. The small difference in the normalization of the IO weights can have a big impact on estimator behavior when weights are highly variable [34], because estimator (2) may produce estimates that fall outside the support of the outcome variable, whereas estimator (3) always produces results that fall in the support of the outcome variable.

5.3 | Doubly robust estimators

In practical applications, background knowledge is typically inadequate to ensure correct specification of the working models for the probability of participation or the expectation of the outcome, and misspecification of these models can lead to estimator inconsistency. We can gain some robustness to misspecification compared to weighting estimators by combining the two models to obtain *doubly robust* estimators that are consistent when either model is correctly specified [35, 36, 37]. Here, we examine three doubly robust estimators that are easy to implement in standard statistical software.

The first doubly robust estimator we consider relies on estimating models for the conditional expectation of the outcome, $g_a(X; \theta)$; the probability of trial participation, $p(X; \beta)$; and (optionally) the probability of treatment among trial participants, $e_a(X; \gamma)$. Predicted values from these models are then combined to obtain the estimator

$$\widehat{\mu}_{\text{DR1}}(a) = \left\{ \sum_{i=1}^{n}(1 - S_i) \right\}^{-1} \sum_{i=1}^{n} \left\{ \widehat{w}_a(X_i, S_i, A_i)\left\{ Y_i - g_a(X_i; \widehat{\theta}) \right\} + (1 - S_i)g_a(X_i; \widehat{\theta}) \right\}, \tag{4}$$

where $\widehat{w}_a(X_i, S_i, A_i)$ was defined in the previous section.

Using the normalized IO weights, an alternative doubly robust estimator is

$$\widehat{\mu}_{\text{DR2}}(a) = \left\{ \sum_{i=1}^{n} \widehat{w}_a(X_i, S_i, A_i) \right\}^{-1} \sum_{i=1}^{n} \widehat{w}_a(X_i, S_i, A_i)\left\{ Y_i - g_a(X_i; \widehat{\theta}) \right\} + \left\{ \sum_{i=1}^{n}(1 - S_i) \right\}^{-1} \sum_{i=1}^{n}(1 - S_i)g_a(X_i; \widehat{\theta}). \tag{5}$$

A third doubly robust estimator involves fitting a model for the outcome conditional on covariates among trial participants, using a weighted regression with the weights $\widehat{w}_a(X_i, S_i, A_i)$ as defined above, and then averaging the predicted values, $g_a(X_i; \widehat{\theta}_w)$ over the covariate distribution of non-participants,

$$\widehat{\mu}_{\text{DR3}}(a) = \left\{ \sum_{i=1}^{n}(1 - S_i) \right\}^{-1} \sum_{i=1}^{n}(1 - S_i)g_a(X_i; \widehat{\theta}_w), \tag{6}$$

where $\widehat{\theta}_w$ is the vector of estimated parameters from the weighted outcome regression. This estimator is doubly robust when the outcome is modeled with a linear exponential family quasi-likelihood [38] and the canonical link function [34, 39].

Doubly robust estimators are consistent for $\mu(a)$ and asymptotically normal when either the model for the probability of participation or the expectation of the outcome is correctly specified [36, 40] (see the Appendix for a formal result). More important for applied work is that doubly robust estimators usually produce estimates that are more precise than those of IO weighting estimators and nearly as precise as those of the outcome model-based estimator, while giving investigators two oportunities for valid inference. In some cases, however, when misspecification of the outcome model is combined with highly variable weights doubly robust estimators can perform worse than non-doubly robust estimators that use the same misspecified outcome model [41, 34].

5.4 | Inference

When using parametric working models, all the estimators described above can be viewed as partial M-estimators [42] and it is possible to employ the usual "sandwich" approach to obtain their sampling variances (e.g., [43, 44]). Inference based on the non-parametric bootstrap [45], however, is easy to obtain and will often be preferred in practice.

6 | SIMULATION STUDY

We conducted proof-of-concept simulation studies to examine the finite-sample performance of different estimators for potential outcome means and treatment effects in the target population. In this Section, we describe simulation studies with continuous outcomes in nested trial designs and present results from select scenarios from those simulations. In the Appendix, we report complete results from simulations with continuous and binary outcomes for nested trial designs, as well as the methods and results of additional simulation studies for non-nested trial designs.

6.1 | Data generation

We run a factorial experiment using 3 trial sample sizes (n_{RCT}) $\times$ 3 total cohort sample sizes (n) $\times$ 2 magnitudes of departure from additive effects in the outcome model $\times$ 2 magnitudes of selection into the trial, resulting in a total of 36 simulation scenarios.

We considered cohorts of $n = 2000$, 5000, or 10,000 individuals with the number of randomized individuals n_{RCT} approximately equal to 200, 500, or 1000. Specifically, we generated baseline covariates for cohorts of 2000, 5000, or 10,000 individuals as $X = (1, X_1, X_2, X_3)^{\top}$, and for $j = 1, 2, 3$, X_j followed an independent standard normal distributions. We then simulated selection into the trial using a binary indicator S with logit $\Pr[S = 1|X] = \beta X$, with $\beta = (\beta_0, \dots, \beta_3)$. We chose β_0 using numerical methods [46], such that for a given cohort sample size n the trial sample size n_{RCT} was 200, 500, or 1000. We examined scenarios with $\beta_1 = 0$, indicating no selection on X_1, and $\beta_1 = 1$, indicating strong selection; in all scenarios we used $\beta_2 = \beta_3 = 1$, indicating strong selection on X_2 and X_3.

We generated potential outcomes as $Y^a = \theta^a X + \epsilon^a$. We used $\theta^1 = (1, \theta_1^1, 1, 1)$, with $\theta_1^1 = 2$ or 1. We used $\theta^0 = (0, 1, 1, 1)$ in all scenarios. We generated the errors ϵ^a using independent standard normal distributions. Note that $\theta_1^1 = 2$ indicates the presence of effect modification on the mean difference scale. For observations "selected" into the trial ($S = 1$) we generated treatment assignment A as a binomial random variable with parameter 0.5 (i.e., we simulated marginally randomized trials). Last, we generated observed outcomes as $Y = AY^1 + (1 - A)Y^0$.

For each simulated dataset, we applied the estimators in equations (1) through (6), and also obtained a trial-only estimator of the treatment effect. All working models required for the different estimators were correctly specified, in the sense that the true models were nested within the parametric working models on which the estimators relied. Specifically, outcome models included main effects for all covariates and were fit separately in each arm; logistic regression models for trial participation and treatment included the main effects of all covariates; all models had intercept terms. We estimated the bias and variance for each estimator over 100,000 runs for each scenario.

6.2 | Simulation results

Tables 2 and 3 summarize simulation results from selected simulation scenarios for continuous, normally distributed outcomes, and linear outcome models in nested trial designs. Additional simulation results are presented in the Appendix, including scenarios with binary outcomes and non-nested trial designs.

When all models were correctly specified, all estimators were approximately unbiased, even with fairly small trial and target population sample sizes. The outcome-model based estimator had the lowest variance, followed closely by the three doubly robust estimators. The IO weighting estimators had substantially larger variance than all other estimators; that variance, though, became smaller with increasing trial sample sizes. When trial sample size was much smaller than the target sample size, in the presence of strong selection on covariates, estimators that used normalized weights (IOW2, DR2, and DR3 in the Tables) had smaller variance compared to estimators that used unnormalized weights (IOW1 and DR1 in the Tables). As expected, in the presence of selection on a covariate that is also an effect modifier, the trial-only estimator gave different results compared to the estimators in equations (1) through (6). Specifically, the trial-only estimator is biased for $E[Y^1 - Y^0 | S = 0]$ when selection into the trial depends on the effect modifier, but is, of course, unbiased for $E[Y^1 - Y^0 | S = 1]$ under very general conditions.

A note about model misspecification in our simulation studies: Even though the simulations did not explicitly examine misspecified models, they nevertheless provide a good illustration of the double robustness property! This counter-intuitive claim is justified because the IO weighting estimators can be viewed as special versions of the doubly robust estimators with grossly misspecified models of the expectation of the outcome; specifically, models that produce predictions $g_a(X; \widehat{\theta})$ identically equal to 0. Similarly, the outcome model-based estimators can be viewed as special versions of the doubly robust estimators with grossly misspecified models for the probability of trial participation; specifically, models that produce probabilities of trial participation $p(X; \widehat{\beta})$ identically equal to 1, so that the weights $\widehat{w}_a(X, S, A)$ are identically equal to 0. Because our IO weighting and outcome regression estimators (with correctly specified probability of participation and outcome models, respectively) are approximately unbiased, viewing these estimators as versions of doubly robust estimators (with grossly misspecified outcome or probability of participation models, respectively), verifies the double robustness property.

TABLE 2 Selected results from the simulation study: bias of estimators across sample sizes, under strong selection on the effect modifier ($\beta_1 = 1$), strong effect modification ($\theta_1^1 - \theta_1^0 = 1$).

n_{RCT}	n	Trial	IOW1	IOW2	OM	DR1	DR2	DR3
200	2000	0.7656	0.0018	0.0593	0.0000	0.0000	0.0003	0.0006
200	5000	0.8401	0.0080	0.0710	-0.0007	0.0002	-0.0002	0.0000
200	10000	0.8855	0.0029	0.0753	-0.0002	-0.0019	-0.0015	-0.0012
500	2000	0.6892	0.0036	0.0235	0.0001	0.0016	0.0013	0.0013
500	5000	0.7654	-0.0003	0.0306	-0.0001	0.0004	0.0002	-0.0004
500	10000	0.8233	0.0033	0.0339	-0.0012	-0.0023	-0.0021	-0.0020
1000	2000	0.6576	-0.0004	0.0081	0.0000	-0.0004	-0.0004	-0.0003
1000	5000	0.7061	0.0033	0.0157	-0.0003	-0.0009	-0.0009	-0.0005
1000	10000	0.7653	-0.0010	0.0160	0.0002	0.0002	0.0002	0.0002

Trial = trial-only, unadjusted estimator; OM = outcome model-based estimator; IOW1 = inverse odds weighting estimator without normalized weights; IOW2 = inverse odds estimator with normalized weights; DR1 = doubly robust estimator without normalized weights; DR2 = doubly robust estimator with normalized weights; DR3 = doubly robust estimator based on weighted multi-variable regression. Additional simulation results are presented in the Appendix.

TABLE 3 Selected results from the simulation study: variance of estimators across sample sizes, under strong selection on the effect modifier ($\beta_1 = 1$) and strong effect modification ($\theta_1^1 - \theta_1^0 = 1$).

n_{RCT}	n	Trial	IOW1	IOW2	OM	DR1	DR2	DR3
200	2000	0.0843	4.4507	1.1256	0.0762	0.2562	0.1501	0.1582
200	5000	0.0905	6.9650	1.2765	0.0819	0.3353	0.1669	0.1791
200	10000	0.0939	6.1245	1.3539	0.0841	0.3321	0.1753	0.1921
500	2000	0.0324	1.1901	0.5278	0.0268	0.0808	0.0610	0.0558
500	5000	0.0336	1.5804	0.6760	0.0294	0.1059	0.0718	0.0662
500	10000	0.0351	1.4593	0.7505	0.0310	0.1113	0.0783	0.0727
1000	2000	0.0166	0.4580	0.2456	0.0127	0.0333	0.0290	0.0258
1000	5000	0.0162	0.6061	0.3558	0.0137	0.0423	0.0346	0.0305
1000	10000	0.0168	0.6964	0.4291	0.0146	0.0503	0.0399	0.0352

Trial = trial-only, unadjusted estimator; OM = outcome model-based estimator; IOW1 = inverse odds weighting estimator without normalized weights; IOW2 = inverse odds weighting estimator with normalized weights; DR1 = doubly robust estimator without normalized weights; DR2 = doubly robust estimator with normalized weights; DR3 = doubly robust estimator based on weighted multivariable regression. Additional simulation results are presented in the Appendix.

6.3 | Code to implement the methods

In the Appendix, we provide R [47] code implementing the methods compared in the simulation study. Specifically, we provide a collection of basic stand-alone functions, one for each estimator in equations (1) through (6), using parametric working models estimated by standard maximum likelihood methods. Readers can modify the functions to incorporate alternative estimation approaches and to obtain bootstrap-based inference. To allow inference with sandwich standard errors, we also provide an implementation of the estimators using the R package geex [48]. Last, we provide Stata code to reproduce the simulation study.

7 | ANALYSES IN THE CORONARY ARTERY SURGERY STUDY

The Coronary Artery Surgery Study (CASS) included a randomized trial nested within a cohort study (i.e., could be viewed as a nested trial design in the terminology of Section 2), comparing coronary artery surgery plus medical therapy (henceforth, "surgery") versus medical therapy alone for patients with chronic coronary artery disease. Of the 2099 eligible patients, 780 consented to randomization and 1319 declined. We excluded six patients for consistency with prior CASS analyses [49, 50] and in accordance with CASS data release notes; in total, we used data from a total of 2093 patients. Details about the design of the CASS are available elsewhere [51, 52]. Here, we focus on estimating the survival probability and treatment effects among eligible non-participants and comparing them against estimates obtained among trial participants.

We implemented the methods described in Section 5 to estimate the 10-year risk (cumulative incidence proportion) of death from any cause in the surgery and medical therapy groups, the risk difference, and the relative risk for the population of eligible patients who did not consent to randomization. Risks are reasonable measures of incidence in

FIGURE 1 Kernel densities for the estimated probability of trial participation for trial participants (left panel, solid line) and non-participants (left panel, dashed line), and the estimated weights for trial participants (right panel). The weights for trial participants are equal to the inverse of the estimated odds of trial participation times the inverse of the estimated probability of receiving the treatment actually received, as defined in Section 5.2.
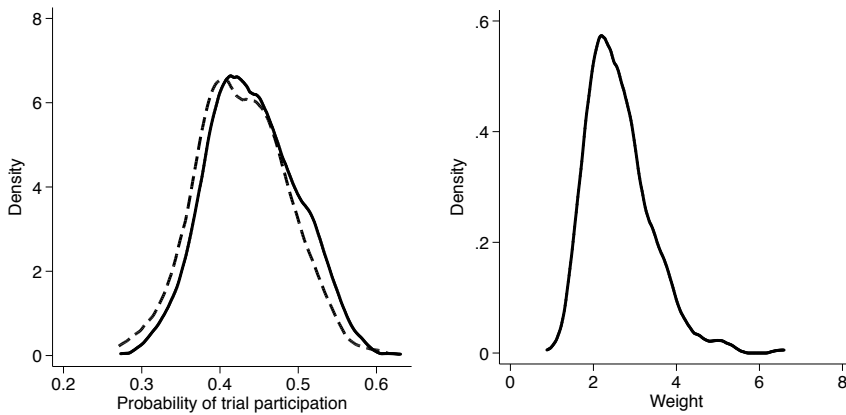
CASS because no patients were censored during the first 10 years of follow-up. The working models for the outcome, the probability of participation in the trial, and the probability of treatment were logistic regression models with the following covariates: age, severity of angina, history of myocardial infarction, percent obstruction of the proximal left anterior descending artery, left ventricular wall motion score, number of diseased vessels, and ejection fraction. We selected variables for inclusion in the models based on a previous analysis of the CASS data [50]; age and ejection fraction were modeled using restricted cubic splines with 5 knots [53]. We used bootstrap re-sampling with 10,000 samples to obtain percentile-based 95% confidence intervals.

Of the 2093 patients in the CASS dataset, 1686 had complete data on all baseline covariates (731 randomized, 368 to surgery and 363 to medical therapy; 955 non-randomized, 430 receiving surgery and 525 medical therapy). For simplicity, we only report analyses restricted to patients with complete data. Table 4 summarizes baseline covariates in trial participants and non-participants (by treatment group). Figure 1 presents the kernel density of the estimated probability of trial participation for trial participants and non-participants and a kernel density of the estimated weights for trial participants. The sample proportion of non-participants divided by the sample average of the IO of trial participation among trial participants was approximately 1.001. We present a comparison of the post-weighting covariate distribution in trial participants against the covariate distribution in non-participants in Section E of the Appendix.

Estimates of the 10-year risk (by treatment group), risk difference, and risk ratio are shown in Table 5. The outcome model-based estimator (OM), the IO weighting estimators (IOW1 and IOW2), and the doubly robust estimators (DR1, DR2, and DR3) produced similar results, suggesting that findings are not driven by model specification decisions [37]. Furthermore, the results from these estimators were similar to those from the trial-only analyses, reflecting the similarity of covariate distributions between randomized and non-randomized individuals. In general, differences will be more striking when the covariate distributions are more different and in the presence of strong effect modification by measured variables. Nevertheless, conducting the analyses we describe in this tutorial is necessary to obtain proper estimates of uncertainty for the potential outcome means or treatment effects in the target population.

TABLE 4 Baseline covariates in the CASS (August 1, 1975 to Dec 31, 1989).

Variable	Levels	Non-randomized group		Randomized group	
		Surgery (430)	Medical therapy (525)	Surgery (368)	Medical therapy (363)
Age, years		51.3 (7.7)	50.6 (7.8)	51.4 (7.2)	50.9 (7.4)
Angina	None	70 (16.3%)	125 (23.8%)	83 (22.6%)	81 (22.3%)
	Present	360 (83.7%)	400 (76.2%)	285 (77.4%)	282 (77.7%)
History of MI	No	194 (45.1%)	212 (40.4%)	159 (43.2%)	135 (37.2%)
	Yes	236 (54.9%)	313 (59.6%)	209 (56.8%)	228 (62.8%)
LAD % obstruction		48.2 (39.8)	31.7 (36.2)	36.4 (38.0)	34.9 (37.0)
Left ventricular score		7.1 (2.7)	7.1 (2.7)	7.4 (2.9)	7.3 (2.8)
Diseased vessels	0	113 (26.3%)	234 (44.6%)	146 (39.7%)	133 (36.6%)
	≥ 1	317 (73.7%)	291 (55.4%)	222 (60.3%)	230 (63.4%)
Ejection fraction, %		60.2 (12.0)	60.1 (12.5)	60.9 (13.0)	59.8 (12.8)

Results presented as mean (SD) for continuous variables and count (%) for discrete variables.
CASS = Coronary Artery Surgery Study; LAD = left anterior descending coronary artery; MI = myocardial infarction; SD = standard deviation.

TABLE 5 Estimated 10-year mortality risk, risk difference, and risk ratio for surgery vs. medical therapy among trial participants and non-participants in the Coronary Artery Surgery Study (August 1, 1975 to Dec 31, 1989).

Estimator	Survival probability (Surgery)	Survival probability (Medical)	Risk difference	Risk ratio
Trial-only	17.4% (13.6%, 21.4%)	20.4% (16.3%, 24.6%)	-3.0% (-8.7%, 2.7%)	0.85 (0.62, 1.15)
OM	18.9% (13.9%, 22.7%)	20.1% (15.9%, 24.5%)	-1.3% (-7.9%, 4.2%)	0.94 (0.65, 1.24)
IOW1	18.2% (13.9%, 22.7%)	20.1% (15.9%, 24.4%)	-1.9% (-7.8%, 4.2%)	0.91 (0.66, 1.24)
IOW2	18.2% (14.6%, 23.5%)	20.1% (16.0%, 24.4%)	-1.9% (-7.2%, 4.8%)	0.91 (0.69, 1.28)
DR1	18.7% (14.5%, 23.3%)	20.1% (16.0%, 24.4%)	-1.4% (-7.3%, 4.7%)	0.93 (0.68, 1.27)
DR2	18.7% (14.5%, 23.3%)	20.1% (16.0%, 24.4%)	-1.4% (-7.3%, 4.7%)	0.93 (0.68, 1.27)
DR3	18.7% (14.4%, 23.2%)	20.0% (15.9%, 24.3%)	-1.4% (-7.3%, 4.6%)	0.93 (0.68, 1.27)

Numbers in parentheses are 95% non-parametric percentile bootstrap confidence intervals based on 10,000 bootstrap samples. Trial = trial-only, unadjusted estimator; OM = outcome model-based estimator; IOW1 = inverse odds weighting estimator without normalized weights; IOW2 = inverse odds weighting estimator with normalized weights; DR1 = doubly robust estimator without normalized weights; DR2 = doubly robust estimator with normalized weights; DR3 = doubly robust estimator based on weighted multi-variable regression; Surgery = coronary artery bypass grafting surgery plus medical therapy; Medical = medical therapy.

8 | PRACTICAL ISSUES

We now discuss practical issues related to lack of exchangeability over S, covariate selection, other aspects of model specification, and positivity violations and highly variable weights, all of which can affect applied analyses.

8.1 | Violations of the exchangeability assumption

All methods described in this tutorial depend critically on the assumption of conditional mean exchangeability over S. When the assumption does not hold, the results produced by the various estimators described in this paper do not have a causal interpretation. Because the mean exchangeability assumption is not testable, it is reasonable to explore how conclusions would change under different degrees of assumption violations. Depending on the available data and background knowledge, various sensitivity analysis methods for missing data or unmeasured confounding can be modified for use in analyses extending inferences to a target population (e.g., [54, 55, 56]). In a companion paper, we have proposed easy to implement estimators for sensitivity analysis that are based on the estimators described in Section 5 and do not require extensive background knowledge about the distribution of unknown or unmeasured variables that are the source of violations of the exchangeability condition [56].

8.2 | Covariate selection

Throughout, we have used X to signify baseline covariates measured both among randomized trial participants and the sample from the target population. In principle, investigators can use any subset of the available covariates that satisfies the assumption of conditional exchangeability over S. When investigators are interested in estimating the potential outcome mean under each treatment (not only the average treatment effect), outcome predictors that are also associated with trial participation should be included in models for the outcome and the probability of participation (or both, when using doubly robust estimators). Including outcome predictors that are not associated with trial participation in models for the expectation of the outcome will often decrease the variance of the outcome model-based and doubly robust estimators; including strong predictors of trial participation that are not associated with the outcome in regressions for the probability of participation will increase the variance of the IO weighting and doubly robust estimators without improving the ability to extend inferences to the target population. When investigators are primarily interested in the average treatment effect (instead of the potential outcome mean under each treatment), only effect modifiers (on the mean difference scale) need to be modeled [12, 8]. Because background knowledge about effect modification is typically limited, even when interest is centered on treatment effect estimation, it is probably best to include as many outcome predictors as possible in regression models for the expectation of the outcome or the probability of trial participation. We followed this strategy in our CASS re-analysis: we selected covariates for "adjustment" based on prior work on outcome modeling and used the same covariates when modeling trial participation, the outcome, and treatment in the trial.

8.3 | Other aspects of model specification

Models for the expectation of the outcome and the probability of trial participation need to approximate the corresponding "true" conditional expectation/probability functions. Covariate balance assessments [57] comparing the post-weighting covariate distribution in trial participants against the (unweighted) covariate distribution in non-participants can provide some guidance for the choice of functional form. When considering parametric models, the methods de-

scribed in [58] can be easily modified to derive the asymptotic bias induced by model misspecification for the estimators described in our paper.

To mitigate model misspecification, investigators can use flexible parametric models (e.g., by including splines for continuous variables or interactions between regressors) or semi- and non-parametric models (e.g., machine learning). When using very flexible models for the outcome or the probability of participation, investigators need to consider the rate of convergence of the working model estimators and its impact on the behavior of potential outcome mean or treatment effect estimators. A detailed discussion of this subtle issue is beyond the scope of the present tutorial (see, [59] for a detailed treatment), but in general doubly robust estimators should be the approach of choice in this context, because the estimators remain $\sqrt{n}$-consistent even when using estimators for the conditional expectation of the outcome or the probability of participation that converge at slower than $\sqrt{n}$-rate [60, 59]. The doubly robust estimators of $\mu(a)$ and the machine learning methods (for estimating the conditional expectation of the outcome or the conditional probability of trial participation) can be combined with cross-fitting methods (a sample splitting approach that reduces overfitting bias) to achieve good behavior under weak assumptions about the underlying models and the properties of the machine learning methods [59]. Also in the spirit of flexible model specification, for outcome model-based or doubly robust estimators, we recommend fitting separate regression models for the outcome in each treatment group in the trial, as we did in the CASS re-analysis (equivalent to fitting a single regression model that includes all possible treatment-covariate interactions).

More broadly, model specification for extending inferences from a trial to a target population involves trading off bias and variance using informal [61] or formal methods (e.g., [62]). When background knowledge suggests that a large number of covariates need to be modeled, formal model specification search methods can be particularly helpful. In our experience, especially when using composite datasets, the variables measured both among trial participants and the sample of the target population are often few and model specification is not a pressing concern (of course, such cases raise concerns about violations of the identifiability conditions and necessitate sensitivity analyses). When richer data are available (e.g., when trials are nested in cohorts of eligible individuals [16]), we may need to employ more sophisticated strategies for model specification search (e.g., [63] provides an overview in the context of causal inference for observational studies, but the same principles apply here).

8.4 | Positivity violations and highly variable weights

To prevent structural violations of the positivity of trial participation assumption, investigators should ensure that the sample from the target population meets the trial eligibility criteria. For example, if the trial restricted enrollment to patients under 85 years of age, it is prudent to apply the same restriction in the sample of patients from the target population. When positivity is violated, IO weighting estimators are inconsistent, whereas outcome model-based and doubly robust estimators rely heavily on the specification of the outcome model (to extrapolate from participants to non-participants) [30]. Empirical (finite-sample) violations of positivity can arise due to chance, particularly when the trial sample size is small or when adjustment for a high-dimensional covariate vector is needed. Empirical violations of positivity increase bias and variance in a way that depends on the particular estimator being used, model specification, and the underlying data generating mechanism.

It is always a good idea to examine the distribution of the estimated probabilities of trial participation, even when using the outcome model-based estimator, because estimated probabilities of trial participation near zero are warning signs for possible positivity violations. Inspection of the estimated probabilities of trial participation can be combined with diagnostics for positivity violations [30]. It is also useful to inspect the distribution of the weights that are used for the IO weighting and doubly robust estimators. By inspecting the distribution of the IO weights, investigators can

identify extreme values and visually assess the spread of the weight distribution. A useful diagnostic is that the sample proportion of non-participants divided by the sample average of the estimated IO of trial participation among trial participants, should be approximately equal to one; or, in symbols

$$\left\{ \sum_{i=1}^{n} (1 - S_i) \right\}^{-1} \sum_{i=1}^{n} S_i \frac{1 - p(X_i; \widehat{\beta})}{p(X_i; \widehat{\beta})} \approx 1.$$

Values different from 1, suggest positivity violations or model misspecification. The rationale for the diagnostic is provided by the identity $E \left[S \frac{\Pr[S = 0|X]}{\Pr[S = 1|X]} \right] = \Pr[S = 0]$.

In applied analyses, we have found that problems with extreme weights can often be addressed by making sensible modeling choices [61] and ensuring that the sample of non-participants is properly selected to avoid violations of the positivity of trial participation assumption. Trimming or truncation of extreme weights may also help, but these strategies shift the causal estimand, which is often undesirable.

In our CASS re-analyses, the estimated probabilities of trial participation were far from zero. Their distribution was similar among trial participants and non-participants (as shown in Figure 1), reflecting the fairly similar observed covariate distribution in trial participants and non-participants and the absence of strong selection into the trial (at least based on available covariates, as shown in Table 4). As noted, the sample proportion of non-participants divided by the sample average of the IO of trial participation among trial participants was approximately 1, providing some reassurance that gross violations of positivity were absent. After weighting, the covariates included in the model for the probability of trial participation were well-balanced (Appendix Section E).

9 | DISCUSSION

In this tutorial, we reviewed methods for extending causal inferences about time-fixed treatments from a randomized trial to a target population of non-participants using baseline covariate data from randomized participants and a sample from the target population, but treatment and outcome data only from the randomized participants. We considered estimation approaches that rely on modeling the probability of trial participation, the expectation of the outcome, or both, and can be implemented easily in all popular statistical software packages.

A major challenge in applying any of the methods discussed in this tutorial is the need to collect adequate covariate information, both from trial participants and non-participants, for the assumption of conditional exchangeability in mean over S to hold. Because this assumption is not testable using the observed data, we have to rely on background substantive knowledge to assess its plausibility. Reasoning about the assumption can be facilitated using directed acyclic graphs, including recent graphical identification algorithms for assessing generalizability/transportability [64, 65, 66]. Because background knowledge is typically incomplete, it is often necessary to conduct sensitivity analyses, to examine how violations of the exchangeability assumption influence study results [67, 68].

Methods related to those discussed in this tutorial have been discussed in a number of recent publications [6, 7, 8, 9, 10, 11, 12, 15, 13, 14]. With few exceptions – such as the careful asymptotic study of an estimator closely related to DR1 in [12], or the targeted maximum likelihood estimators in [13] – prior work has focused on weighting [6, 11, 14] or stratification-based methods [8, 9, 10] that only rely on the probability of trial participation. Theoretical arguments, our simulation results, and practical experience suggest that methods that combine modeling the probability of trial participation with modeling the expectation of the outcome are most promising for applied work for two reasons: first, the double robustness property in effect gives investigators two opportunities for approximately correct inference [37]; second, doubly robust estimators often produce estimates that are more precise than those from methods

that exclusively rely on modeling the probability of trial participation, even when the outcome model is misspecified [37, 34, 69, 40]. In our proof-of-concept simulation studies, which used correctly specified parametric working models with few covariates, all estimators performed reasonably well in terms of bias. Interestingly, the two IO weighting estimators had very different finite-sample performance in the presence of strong selection. Based on this observation, we recommend avoiding weighted estimators that do not use normalized weights.

When data are available on numerous baseline covariates, many of which are continuous, correct specification of parametric models for the probability of trial participation or expectation of the outcome will be impossible. Future research should address estimation using more flexible models (e.g., non-parametric or semi-parametric regression, machine learning) to mitigate model misspecification [59]. Further research is also needed to study the behavior of different estimators under misspecification and to develop alternatives that are more robust to misspecification of the outcome model (e.g., along the lines suggested in [70, 71]). Lastly, throughout this tutorial, we have assumed perfect adherence to treatment in the randomized trial, no missing outcome data, and no measurement error. In practice, adherence is often imperfect, outcomes are missing (e.g., due to right censoring in failure-time analyses), and measurement error is a concern (e.g., differential measurement error in effect modifiers when using composite datasets). Established methods to address these issues in the trial data can be combined with the methods described in this tutorial in a modular fashion. For example, adjustment for imperfect adherence via inverse probability of treatment weighting can be combined with the methods reviewed in this tutorial [29]. Future work should assess the properties of such combined procedures and evaluate them in practical applications.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST

The authors have no conflicts of interest to report.

REFERENCES

[1] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". The Lancet 2005;365(9453):82–93.

[2] Evans A, Kalra L. Are the results of randomized controlled trials on anticoagulation in patients with atrial fibrillation generalizable to clinical practice? Archives of Internal Medicine 2001;161(11):1443–1447.

[3] Elting LS, Cooksley C, Bekele BN, Frumovitz M, Avritscher EB, Sun C, et al. Generalizability of cancer clinical trial results. Cancer 2006;106(11):2452–2458.

[4] Steg PG, López-Sendón J, de Sa EL, Goodman SG, Gore JM, Anderson FA, et al. External validity of clinical trials in acute myocardial infarction. Archives of Internal Medicine 2007;167(1):68–73.

[5] Dans AL, Dans LF, Guyatt GH, Richardson S, Group EBMW, et al. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. JAMA 1998;279(7):545–549.

[6] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. American Journal of Epidemiology 2010;172(1):107–115.

[7] Kaizar EE. Estimating treatment effect via simple cross design synthesis. Statistics in Medicine 2011;30(25):2986–3009.

[8] O'Muircheartaigh C, Hedges LV. Generalizing from unrepresentative experiments: a stratified propensity score approach. Journal of the Royal Statistical Society Series C (Applied Statistics) 2014;63(2):195–210.

[9] Tipton E. Improving Generalizations From Experiments Using Propensity Score Subclassification Assumptions, Properties, and Contexts. Journal of Educational and Behavioral Statistics 2012;38(3):239–266.

[10] Tipton E, Hedges L, Vaden-Kiernan M, Borman G, Sullivan K, Caverly S. Sample selection in randomized experiments: A new method using propensity score stratified sampling. Journal of Research on Educational Effectiveness 2014;7(1):114–135.

[11] Hartman E, Grieve R, Ramsahai R, Sekhon JS. From SATE to PATT: combining experimental with observational studies to estimate population treatment effects. Journal of the Royal Statistical Society Series A (Statistics in Society) 2013;10:1111.

[12] Zhang Z, Nie L, Soon G, Hu Z. New methods for treatment effect calibration, with applications to non-inferiority trials. Biometrics 2016;72(1):20–29.

[13] Rudolph KE, van der Laan MJ. Robust estimation of encouragement design intervention effects transported across sites. Journal of the Royal Statistical Society Series B (Statistical Methodology) 2017;79(5):1509–1525.

[14] Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. American Journal of Epidemiology 2017;186(8):1010–1014.

[15] Buchanan AL, Hudgens MG, Cole SR, Mollan KR, Sax PE, Daar ES, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. Journal of the Royal Statistical Society Series A (Statistics in Society) 2018;181(4):1193–1209.

[16] Dahabreh IJ, Robertson SE, Tchetgen Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 2018;.

[17] Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. European Journal of Epidemiology 2019;p. 1–4.

[18] Hernán M. Discussion of "Perils and potentials of self-selected entry to epidemiological studies and surveys". Journal of the Royal Statistical Society Series A (Statistics in Society) 2016;179(2):346–347.

[19] Dahabreh IJ, Haneuse SJP, Robins JM, Robertson SE, Buchanan AL, Stuart EA, et al. Study designs for extending causal inferences from a randomized trial to a target population. arXiv preprint arXiv:190507764 2019;.

[20] Olschewski M, Scheurlen H. Comprehensive Cohort Study: an alternative to randomized consent design in a breast preservation trial. Methods of Information in Medicine 1985;24:131–134.

[21] Fiore LD, Lavori PW. Integrating randomized comparative effectiveness research with patient care. New England Journal of Medicine 2016;374(22):2152–2158.

[22] Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. American Journal of Epidemiology 2003;158(9):915–920.

[23] Splawa-Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. [Translated from Splawa-Neyman, J (1923) in Roczniki Nauk Rolniczych Tom X, 1–51]. Statistical Science 1990;5(4):465–472.

[24] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974;66(5):688.

[25] Hernán MA, Robins JM. Causal inference (forthcoming). Boca Raton, FL: Chapman & Hall/CRC; 2020.

[26] Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International Journal of Epidemiology 2016;45(6):2184–2193.

[27] Dahabreh IJ, Trikalinos TA, Kent DM, Schmid CH. Heterogeneity of Treatment Effects. Methods in Comparative Effectiveness Research 2017;p. 227.

[28] Dahabreh IJ, Hernán MA, Robertson SE, Buchanan A, Steingrimsson JA. Generalizing trial findings in nested trial designs with sub-sampling of non-randomized individuals. arXiv preprint arXiv:190206080 2019;.

[29] Dahabreh IJ, Robins JM, Haneuse SJP, Hernán MA. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. arXiv preprint arXiv:190610792 2019;.

[30] Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Statistical Methods in Medical Research 2012;21(1):31–54.

[31] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. Mathematical Modelling 1986;7(9):1393–1512.

[32] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 1952;47(260):663–685.

[33] Hájek J. Comment on "An essay on the logical foundations of survey sampling by D. Basu". In: Godambe VP, Sprott DA, editors. Foundations of statistical inference New York City, NY: Holt, Rinehart, and Winston; 1971.p. 236.

[34] Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. Statistical Science 2007;22(4):544–559.

[35] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 1994;89(427):846–866.

[36] Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 1995;90(429):122–129.

[37] Robins JM, Rotnitzky A. Comments. Statistica Sinica 2001;11(4):920–936.

[38] Gourieroux C, Monfort A, Trognon A. Pseudo maximum likelihood methods: Theory. Econometrica: Journal of the Econometric Society 1984;p. 681–700.

[39] Wooldridge JM. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics 2007;141(2):1281–1301.

[40] Tsiatis A. Semiparametric theory and missing data. Springer Science & Business Media; 2007.

[41] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 2007;p. 523–539.

[42] Stefanski LA, Boos DD. The calculus of M-estimation. The American Statistician 2002;56(1):29–38.

[43] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 2004;23(19):2937–2960.

[44] Chen Z, Kaizar E. On variance estimation for generalizing from a trial to a target population. arXiv preprint arXiv:170407789 2017;.

[45] Efron B, Tibshirani RJ. An introduction to the bootstrap, vol. 57 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC; 1994.

[46] Austin PC, Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. Communications in Statistics - Simulation and Computation 2008;37(6):1039–1051.

[47] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2015, `https://www.R-project.org`.

[48] Saul BC, Hudgens MG. The Calculus of M-estimation in R with geex. arXiv preprint arXiv:170901413 2017;.

[49] Chaitman BR, Ryan TJ, Kronmal RA, Foster ED, Frommer PL, Killip T, et al. Coronary Artery Surgery Study (CASS): comparability of 10 year survival in randomized and randomizable patients. Journal of the American College of Cardiology 1990;16(5):1071–1078.

[50] Olschewski M, Schumacher M, Davis KB. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. Controlled Clinical Trials 1992;13(3):226–239.

[51] William J, Russell R, Nicholas T, et al. Coronary Artery Surgery study (CASS): a randomized trial of coronary artery bypass surgery. Circulation 1983;68(5):939–950.

[52] Investigators CP. Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery: comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria. Journal of the American College of Cardiology 1984;3(1):114–128.

[53] Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.

[54] Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA, et al. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. The Annals of Applied Statistics 2017;11(1):225–247.

[55] Nguyen TQ, Ackerman B, Schmid I, Cole SR, Stuart EA. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. PloS one 2018;13(12):e0208795.

[56] Dahabreh IJ, Robins JM, Haneuse SJP, Saeed I, Robertson SE, Stuart EA, et al. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. arXiv preprint arXiv:190510684 2019;.

[57] Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology 2001;2(3-4):169–188.

[58] Waernbaum I, Pazzagli L. Model misspecification and bias for inverse probability weighting and doubly robust estimators. arXiv preprint arXiv:171109388 2017;.

[59] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 2018;21(1):C1–C68.

[60] Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine 1997;16(3):285–319.

[61] Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology 2008;168(6):656–664.

[62] Brookhart MA, Van Der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Computational statistics & data analysis 2006;50(2):475–498.

[63] Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Statistical Methods in Medical Research 2012;21(1):7–30.

[64] Bareinboim E, Pearl J. Meta-transportability of causal effects: A formal approach. In: Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS); 2013. p. 135–143.

[65] Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Statistical Science 2014;29(4):579–595.

[66] Bareinboim E, Pearl J. Transportability of Causal Effects: Completeness Results. In: AAAI; 2012. p. 698–704.

[67] Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association 1998;93(444):1321–1339.

[68] Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Statistical models in epidemiology, the environment, and clinical trials Springer; 2000.p. 1–94.

[69] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61(4):962–973.

[70] Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 2009;96(3):723–734.

[71] Vermeulen K, Vansteelandt S. Bias-reduced doubly robust estimation. Journal of the American Statistical Association 2015;110(511):1024–1036.