HHS Public Access
Author manuscript
Biometrics. Author manuscript; available in PMC 2024 March 14.

Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals

Issa J. Dahabreh[1,2,3,*], **Sarah E. Robertson**[1], **Eric J. Tchetgen Tchetgen**[4], **Elizabeth A. Stuart**[5], **Miguel A. Hernán**[3,6,7]

[1]Center for Evidence Synthesis in Health, Brown University School of Public Health, Providence, RI, U.S.A.

[2]Departments of Health Services, Policy & Practice and Epidemiology, Brown University, Providence, RI, U.S.A.

[3]Department of Epidemiology, Harvard-T.H. Chan School of Public Health, Boston, MA, U.S.A.

[4]Department of Statistics, Wharton Business School, University of Pennsylvania, PA, U.S.A.

[5]Departments of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, U.S.A.

[6]Department of Biostatistics, Harvard-T.H. Chan School of Public Health, Boston, MA, U.S.A.

[7]Harvard-MIT Division of Health Sciences and Technology, Boston, MA, U.S.A.

Summary:

We consider methods for causal inference in randomized trials nested within cohorts of trial-eligible individuals, including those who are not randomized. We show how baseline covariate data from the entire cohort, and treatment and outcome data only from randomized individuals, can be used to identify potential (counterfactual) outcome means and average treatment effects in the target population of all eligible individuals. We review identifiability conditions, propose estimators, and assess the estimators' finite-sample performance in simulation studies. As an illustration, we apply the estimators in a trial nested within a cohort of trial-eligible individuals to compare coronary artery bypass grafting surgery plus medical therapy versus medical therapy alone for chronic coronary artery disease.

Keywords

causal inference; clinical trials; double robustness; generalizability; observational studies; transportability

---

[*] issa_dahabreh@brown.edu .

1. Introduction

When effect modifiers influence who participates in randomized trials, causal inferences from *randomized individuals* need to be generalized (extended) to *the population of all trial-eligible individuals* (Rothwell, 2005). The need to extend trial findings arises naturally when trials are nested in cohort studies that collect baseline covariate data from all eligible individuals, including those who are not randomized. In this setting, treatment and follow-up data from non-randomized individuals might be unavailable or unreliable. For example, investigators might wish to use treatment and outcome information only from randomized individuals to avoid confounding of the treatment effect among non-randomized individuals (Torgerson and Sibbald, 1998; Silverman and Altman, 1996).

In this paper, we build on the emerging literature on "generalizing" and "transporting" inferences from randomized trials to a target population (Cole and Stuart, 2010; Tipton, 2012; O'Muircheartaigh and Hedges, 2014; Hartman et al., 2013; Rudolph and van der Laan, 2017; Zhang et al., 2016) to show how data from randomized trials nested within cohorts of eligible individuals can be used to generalize inferences from randomized individuals the target population of all trial-eligible individuals. We review identifiability conditions and propose estimators for the potential (counterfactual) outcome means and average treatment effects in the target population. We assess the finite-sample performance of different estimators in simulation studies. Lastly, we illustrate the application of the estimators in the Coronary Artery Surgery Study (CASS), a randomized trial nested within a cohort of trial-eligible individuals to compare coronary artery bypass grafting surgery plus medical therapy versus medical therapy alone for chronic coronary artery disease.

2. Targets of inference

Consider a trial nested in a cohort of trial-eligible individuals and let $\mathscr{A}$ be the set of treatments evaluated in the randomized trial. For each treatment $a \in \mathscr{A}$, we use the random variable Y^a to denote the potential (counterfactual) outcome under intervention to receive treatment a (Splawa-Neyman, 1990; Rubin, 1974; Robins and Greenland, 2000). We only consider a finite number of distinct treatments (e.g., comparisons of treatment vs. control, or comparisons between two or more active treatments).

Baseline covariate information is collected from all cohort members, but treatment and outcome information is only collected (or only deemed reliable) from randomized individuals. We model the data as independent and identically distributed realizations of the random tuple $(X_i, S_i, S_i \times A_i, S_i \times Y_i)$, $i = 1, \ldots, n_{S=1}, n_{S=1} + 1, \ldots, n_{S=1} + n_{S=0} = n$, where $n_{S=1}$ is the number of randomized individuals, $n_{S=0}$ is the number of non-randomized individuals, and n is the total number of trial-eligible individuals; S is the indicator for being randomized ($S = 1$ for randomized individuals; $S = 0$ for non-randomized trial-eligible individuals); X is a vector of baseline covariates; A is the (randomized) treatment assignment indicator; and Y is the observed outcome. An example data structure for binary treatment A, along with the potential outcomes, is depicted in Table 1 (throughout, we use uppercase letters to denote random variables and lowercase letters to denote realizations). For simplicity, we assume that full adherence to assigned treatment, absence of measurement error, and no dropout in

the trial. Though the methods we propose can be extended to address these issues, we do not pursue these extensions here to simplify exposition and maintain focus on selective trial participation. Extensions of our results will be considered in future work.

We are interested in using the data to draw causal inferences about all trial-eligible individuals. Key targets of inference are the potential outcome means $E[Y^a]$, for each $a \in \mathcal{A}$, and the average treatment effects $E[Y^a - Y^{a'}]$, for any pair of treatments $a, a' \in \mathcal{A}$. In general, $E[Y^a - Y^{a'}] \neq E[Y^a - Y^{a'} \mid S = 1]$; in our setup, differences arise when effects are heterogeneous over baseline covariates that are not equidistributed among randomized and non-randomized individuals (Dahabreh et al., 2016).

3. Identifiability conditions for potential outcome means

The following conditions are sufficient to identify the potential outcome means $E[Y^a]$ for $a \in \mathcal{A}$: *(I) Consistency of potential outcomes:* for individuals who receive treatment $A = a$, the observed outcome equals the potential outcome under treatment a, that is, if $A_i = a$, then $Y_i = Y_i^a$, for every $a \in \mathcal{A}$. *(II) Mean exchangeability in the trial:* $E[Y^a \mid X, S = 1, A = a] = E[Y^a \mid X, S = 1]$ for every $a \in \mathcal{A}$. *(III) Positivity of treatment assignment probability in the trial:* $\Pr[A = a \mid X = x, S = 1] > 0$ for every $a \in \mathcal{A}$ and every x such that $f_{X \mid S}(x \mid S = 1) > 0$. *(IV) Mean generalizability (i.e., mean exchangeability over S):* $E[Y^a \mid X, S = 1] = E[Y^a \mid X]$ for every $a \in \mathcal{A}$. The mean generalizability condition is implied by, but does not imply, the often invoked (and stronger) condition of generalizability in distribution, $Y^a \perp\!\!\!\perp S \mid X$. *(V) Positivity of trial participation:* $\Pr[S = 1 \mid X = x] > 0$ for every x such that $f_X(x) > 0$.

Conditions *I* through *III* are expected to hold for well-defined interventions compared in randomized trials (Hernán and Robins, 2018). Of note, implicit in our notation is an assumption that the invitation to participate in the trial and trial participation itself do not affect the outcome except through treatment assignment – this assumption is often plausible in pragmatic randomized trials (Ford and Norrie, 2016). Conditions *IV* and *V* are needed to extend inferences about the potential outcome means (or the average treatment effect) from randomized individuals to the target population of all trial-eligible individuals.

As is common in the causal inference literature, we use the term "exchangeability" to mean that two or more groups are expected to have the same outcome functionals had they received the same treatment (Greenland and Robins, 1986). This notion differs from other uses of the term in statistics. Furthermore, in this section, we used the term "consistency" for the condition linking potential (counterfactual) and observed outcomes; later, we will use the same term to refer to the property of an estimator that converges to its estimand; the intended meaning should be clear from context.

4. Identification

Under the identifiability conditions listed in the previous section, the conditional potential outcome mean in the target population of all trial eligible individuals for each treatment $a \in \mathscr{A}$ can be identified,

$$\mathrm{E}\big[Y^a \mid X\big] = \mathrm{E}\big[Y^a \mid X, S = 1\big] = \mathrm{E}\big[Y^a \mid X, S = 1, A = a\big] = \mathrm{E}\big[Y \mid X, S = 1, A = a\big].$$

It follows that we can identify the potential outcome means,

$$\mathrm{E}\big[Y^a\big] = \mathrm{E}_X\big[\mathrm{E}\big[Y \mid X, S = 1, A = a\big]\big], \text{ for each } a \in \mathscr{A},$$

where the subscript X denotes expectation with respect to the target population distribution of X. Using this result, we can identify average treatment effects by differencing,

$$\mathrm{E}[Y^a - Y^{a'}] = \mathrm{E}_X[\mathrm{E}[Y \mid X, S = 1, A = a]] - \mathrm{E}_X[\mathrm{E}[Y \mid X, S = 1, A = a']].$$

Furthermore, we can identify any other measure of effect defined in terms of the potential outcome means (Hernán, 2004); for example, for binary Y, we can identify the causal risk ratio comparing treatments a and a' in the target population,

$$\frac{\Pr[Y^a = 1]}{\Pr[Y^{a'} = 1]} = \frac{\mathrm{E}[Y^a]}{\mathrm{E}[Y^{a'}]} = \frac{\mathrm{E}_X[\mathrm{E}[Y \mid X, S = 1, A = a]]}{\mathrm{E}_X[\mathrm{E}[Y \mid X, S = 1, A = a']]}.$$

In the Appendix, we show that identification of average treatment effects is possible under weaker identifiability conditions, which do not, however, suffice to identify the potential outcome means. Because these potential outcome means are of inherent scientific and policy interest in most applications, in the remainder of the paper we focus on estimating them.

5. Estimation

We now discuss estimators of the functional $\mu(a) \equiv \mathrm{E}_X[\mathrm{E}[Y \mid X, S = 1, A = a]]$. Specifically, we consider (1) *outcome model-based* estimators that rely on modeling the expectation of the outcome; (2) *inverse probability (IP) weighting* estimators that rely on modeling the probability of participation in the trial; and (3) *augmented IP weighting* estimators that rely on modeling both the expectation of the outcome and the probability of participation in the trial. Hereafter, "convergence" and the symbol "→" denote convergence in probability; estimators that converge to their corresponding estimands are termed "consistent" (see the Web Appendix for additional information).

5.1 Outcome model-based estimator

The outcome model-based estimator is obtained by fitting a conditional outcome mean model among trial participants and marginilizing over the empirical covariate distribution of all trial-eligible individuals (Robins, 1986),

$$\widehat{\mu}_{\mathrm{OM}}(a) = \frac{1}{n} \sum_{i=1}^{n} \widehat{g}_a(X_i),$$

(1)

where $\widehat{g}_a(X)$ is an estimator of $\mathrm{E}[Y \mid X, S = 1, A = a]$ for $a \in \mathscr{A}$. Typically, we posit a parametric outcome model $g_a(X; \theta)$ with finite-dimensional parameter θ for each treatment a, and estimate the model parameters by standard methods. In applications, we recommend fitting separate outcome models among the treated and untreated randomized individuals to better capture effect modification over baseline covariates. When the outcome models are correctly specified, $g_a(X; \widehat{\theta}) \to \mathrm{E}[Y \mid X, S = 1, A = a]$ and $\widehat{\mu}_{\mathrm{OM}}(a) \to \mu(a)$ for each $a \in \mathscr{A}$.

5.2 IP weighting estimators

An alternative approach for generalizing inferences from randomized individuals to the target population relies on estimating the probability of trial participation followed by IP weighting, an approach related to the analysis of sampling surveys (Horvitz and Thompson, 1952) and to IP weighting methods for confounding control in observational studies (Robins, 1999). Specifically, we can use the IP weighting estimator

$$\widehat{\mu}_{\mathrm{IPW1}}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{S_i I(A_i = a) Y_i}{\widehat{p}(X_i) \widehat{e}_a(X_i)},$$

(2)

where $\widehat{p}(X)$ is an estimator for $\Pr[S = 1 \mid X]$ and $\widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, S = 1]$ for $a \in \mathscr{A}$. When trials are nested within cohorts of eligible individuals, we do not know the probability of trial participation, but we can estimate it, typically using a parametric model $p(X; \beta)$. The probability of treatment in the trial $\Pr[A = a \mid X, S = 1]$ is under the control of the investigators and does not need to be estimated. Nevertheless, estimating the probability of treatment among randomized individuals, say using a parametric model $e_a(X; \gamma)$, can improve efficiency in finite samples. Heuristically, modeling the probability of treatment is beneficial because it addresses random imbalances in baseline covariates among randomized individuals, provided that we properly account for the estimation of γ (Hahn, 1998; Lunceford and Davidian, 2004, Williamson et al., 2014). When the model for the probability of participation is correctly specified, and given that the model for the probability of treatment among randomized individuals is always correctly specified, $p(X; \widehat{\beta}) e_a(X; \widehat{\gamma}) \to \Pr[S = 1 \mid X] \Pr[A = a \mid X, S = 1]$ and $\widehat{\mu}_{\mathrm{IPW1}}(a) \to \mu(a)$ for every $a \in \mathscr{A}$.

The estimator in (2) is unbounded, in the sense that it can produce point estimates that fall outside the support of Y, particularly in the presence of extreme weights (Robins et al., 2007). Using the identity

$$\mathrm{E}\left[\frac{S I(A = a)}{\Pr[S = 1 \mid X] \Pr[A = a \mid X, S = 1]} \right] = 1,$$

(3)

which holds when positivity conditions *III* and *V* hold, we can construct a bounded IP weighting estimator by normalizing the weights to sum to 1 (Hájek, 1971),

$$\widehat{\mu}_{\text{IPW2}}(a) = \left\{ \sum_{i=1}^{n} \frac{S_i I(A_i = a)}{\widehat{p}(X_i)\widehat{e}_a(X_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{S_i I(A_i = a)Y_i}{\widehat{p}(X_i)\widehat{e}_a(X_i)}.$$

(4)

Even when not intending to use IP weighting estimators, analysts should inspect the distribution of the estimated probabilities of trial participation to examine overlap between randomized and non-randomized individuals. Furthermore, analysts should examine whether the sample analog of (3) is approximately satisfied, that is,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{S_i I(A_i = a)}{\widehat{p}(X_i)\widehat{e}_a(X_i)} \approx 1.$$

5.3 Augmented IP weighting estimators

We can combine modeling the probability of trial participation with modeling the expectation of the outcome among randomized individuals to obtain more efficient (augmented) IP weighting estimators (Robins et al., 1994; Bang and Robins, 2005). Augmented IP weighting estimators are also *doubly robust* in the sense that they are consistent and asymptotically normal when either the model for the probability of participation or the model for the expectation of the outcome is correctly specified. Because background knowledge is often inadequate to correctly specify the outcome model, the improved efficiency of augmented IP weighting estimators is often the primary motivation for using them (Tan, 2007), provided the model for the probability of trial participation can be (approximately) correctly specified.

The theory of augmented IP weighting estimation is extensive and multiple estimators, with different behavior in finite samples, are doubly robust (Robins et al., 1994, 2007). Here, we examine three estimators that are easy to implement in standard statistical packages.

We begin by considering the augmented IP weighting estimator

$$\widehat{\mu}_{\text{AIPW1}}(a) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{S_i I(A_i = a)}{\widehat{p}(X_i)\widehat{e}_a(X_i)} \{Y_i - \widehat{g}_a(X_i)\} + \widehat{g}_a(X_i) \right\},$$

(5)

with $\widehat{p}(X)$, $\widehat{e}_a(X)$, and $\widehat{g}_a(X)$ as defined above.

We can normalize the weights, as we did for the IP weighting estimators, to obtain

$$\hat{\mu}_{\text{APP2}}(a) = \left\{ \sum_{i=1}^{n} \frac{S_i I(A_i = a)}{\hat{p}(X_i)\hat{e}_a(X_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{S_i I(A_i = a)}{\hat{p}(X_i)\hat{e}_a(X_i)} \{Y_i - \hat{g}_a(X_i)\} + \frac{1}{n}\sum_{i=1}^{n} \hat{g}_a(X_i).$$

(6)

Alternatively, we can obtain a bounded regression-based augmented IP weighting estimator by fitting an IP weighted parametric multi-variable regression model for the outcome among randomized individuals and then standardizing the predictions over the empirical covariate distribution of all trial-eligible individuals,

$$\hat{\mu}_{\text{AIPW3}}(a) = \frac{1}{n}\sum_{i=1}^{n} g_a(X_i; \tilde{\theta}),$$

(7)

where $\tilde{\theta}$ is the vector of estimated parameters from the IP weighted outcome regression. This estimator is doubly robust when the outcome is modeled with a linear exponential family quasi-likelihood (Gourieroux et al., 1984) and the canonical link function (Robins et al., 2007; Wooldridge, 2007).

In the Web Appendix we show that the estimator in (5) is the one-step in-sample estimator suggested by the influence function for $\mu(a)$ and argue that it is locally efficient (Robins et al., 1994; Robins and Rotnitzky, 1995). The estimators in (6) and (7) are asymptotically equivalent to the estimator in (5). In finite samples, augmented IP weighting estimators will tend to produce more precise results than non-augmented IP weighting estimators, sometimes strikingly so. When using any augmented IP weighting estimator, as for the outcome model-based estimator in (1), we recommend fitting separate outcome models in each treatment group in the randomized trial.

6. Simulation studies

We conducted simulation studies to compare the finite-sample performance of different estimators, for binary and continuous outcomes. Details about the simulation study methods and code to replicate the analyses are provided in the Web Appendix; the simulation results are summarized in Appendix Tables A.2 through A.19. In brief, our numerical studies confirmed that, when all models were correctly specified, all estimators were approximately unbiased even with small numbers of randomized individuals and small total cohort sample sizes (when one model was incorrectly specified, the augmented IP weighting estimators were also approximately unbiased; results not shown). The outcome-model based estimator had the lowest variance, followed closely by the two doubly robust estimators; IP weighting estimators had substantially larger variance than all other estimators, especially when using non-normalized weights.

7. The Coronary Artery Surgery Study (CASS)

7.1 CASS design and data

CASS was a comprehensive cohort study that compared coronary artery bypass grafting surgery plus medical therapy (henceforth, "surgery") versus medical therapy alone for individuals with chronic coronary artery disease; details about the design of CASS are available elsewhere (William et al., 1983; Investigators, 1984). In brief, individuals undergoing angiography in 11 participating institutions were screened for eligibility and the 2099 trial-eligible individuals who met the study criteria were either randomized to surgery or medical therapy (780 individuals), or included in an observational study (1319 individuals). We excluded 6 individuals for consistency with prior CASS analyses and in accordance with CASS data release notes; in total we used data from 2093 individuals (778 randomized; 1315 non-randomized). Baseline covariates were collected from randomized and non-randomized individuals in an identical manner. No randomized individuals were lost to follow-up in the first 10 years of the study; we did not use information on adherence among randomized individuals, in effect assuming that the non-adherence would be similar among all eligible individuals, so that intention-to-treat effects are transportable.

7.2 Statistical analysis

Estimands.—We estimated the 10-year mortality risk under surgery and medical therapy, and the risk difference and risk ratio comparing the treatments for the target population of all trial-eligible individuals.

Model specification.—We fit logistic regression models for the probability of participation in the trial, the probability of treatment among randomized individuals, and the probability of the outcome (in each treatment arm) with the following covariates: age, severity of angina, history of previous myocardial infarction, percent obstruction of the proximal left anterior descending artery, left ventricular wall motion score, number of diseased vessels, and ejection fraction. We chose these variables based on a previous analysis of the same data (Olschewski et al., 1992) and did not perform any model selection.

Missing baseline covariate data.—Of the 2093 trial-eligible individuals, 1686 had complete data on all baseline covariates (731 randomized, 368 in the surgery group and 363 in the medical therapy group; 955 non-randomized). For simplicity, in the main text we only report analyses restricted to individuals with complete data. To examine whether missing data influenced our results, we repeated our analyses using (1) multiple imputation with different models for the missing data conditional on the observed data (multivariate normal multiple imputation and multiple imputation with chained equations) (Little and Rubin, 2014) and (2) IP of missingness weighting for non-monotone missing data (Sun and Tchetgen Tchetgen, 2018).

Inference.—For all analyses, we used bootstrap resampling (with 10,000 samples) to obtain normal distribution-based 95% confidence intervals (results using percentile intervals were very similar and are not shown).

7.3 Results

Baseline characteristics, overlap, and balance.—Web Appendix Table A.20 summarizes baseline covariate information for randomized and non-randomized individuals. The left panel of Figure 1 presents kernel densities of the estimated probabilities of trial participation for randomized and non-randomized individuals; there was good overlap and the smallest estimated probability for trial participation for randomized individuals was approximately 0.273. The right panel of Figure 1 presents the kernel density of the estimated weights among randomized individuals, obtained as the inverse of the estimated probability of trial participation times the inverse of the estimated probability of receiving the treatment actually received among randomized individuals. The sample average of the estimated weights was approximately 1.001, both in the surgery and medical therapy groups; the largest estimated weight was less than 10. Taken together, these results suggest that the observed covariate distributions of randomized and non-randomized individuals had sufficient overlap for attempting to extend inferences from the trial to the target population of all eligible individuals. Baseline covariates in randomized and non-randomized individuals were balanced after IP weighting (Web Appendix Table A.21).

Treatment-specific risks and treatment effects.—Estimates of the 10-year mortality risk and treatment effects at 10 years are shown in Table 2. All generalizability methods produced similar results: the mortality risk in the target population was about 18% for surgery and 20% for medical therapy, corresponding to a risk difference of about 2% and a risk ratio of about 0.9, in favor of surgery. Because different methods rely on different parametric models, agreement across their point estimates suggests that inference is not driven by model specification. The similar confidence interval widths produced by different generalizability estimators reflect the binary nature of the outcome as well as the absence of strong selection on measured baseline covariates in this particular application. The mortality risk among randomized individuals was about 17.3% for surgery and 20.9% for medical therapy, corresponding to a risk difference of about 3.5% in favor of surgery. The closeness of estimates from the randomized trial and our generalizability analyses probably reflects the absence of strong selection on measured baseline covariates in this application.

Using treatment and outcome data on non-randomized individuals to evaluate the generalizability analyses.—In CASS, data on treatment and outcome were collected among non-randomized individuals, even though such data are not necessary for identification under the conditions in Section 3. To evaluate whether outcome models built among randomized individuals were reasonable, we compared the empirical mortality risk among non-randomized individuals who received treatment *a* against the average outcome predictions for the same group of individuals, using models estimated among randomized individuals. The similarity of the empirical risk in non-randomized individuals to the average of the predictions provides some reassurance that the models were not grossly inappropriate for generalizing trial results: for surgery, the empirical risk was 18.6% and the average of the predictions was 19.8%; for medical therapy, the empirical risk was 19.1% and the average of the predictions was 18.9%.

The comparison in the preceding paragraph is only indirectly relevant to the goal of generalizing inferences to the target population of all trial-eligible individuals; after all, the comparison is conditional on treatment received among non-randomized individuals. Another way to evaluate the generalizability analyses, is to compare them against an observational analysis that uses all the data (regardless of S). The validity of such observational analysis requires exchangeability of treatment groups, $E[Y^a \mid X, S, A = a] = E[Y^a \mid X, S]$, and positivity, $\Pr[A = a \mid X, S] > 0$, for $a \in \{0, 1\}$, but does not require identifiability conditions *IV* or *V* (which are necessary to endow the generalizability estimators with a causal interpretation). Using an IP weighting regression estimator for confounding control in the entire CASS dataset (Robins et al., 2007), the potential outcome mean estimates were 17.9% and 19.6% for surgery and medical therapy, respectively; these estimates are similar to those from the generalizability analyses. Because the observational and generalizability analyses rely on different identifiability conditions, agreement between them provides some mutual support (but does not establish the validity of either).

Comparison of methods for handling missing data.—Web Appendix A.22 presents descriptive statistics for missing covariate data; the missing data pattern for baseline covariates was non-monotone. The point estimates from analyses using multiple imputation or IP weighting for missing baseline covariates were nearly identical to those of the complete case analyses and only slightly more precise (Web Appendix Tables A.23 through A.25).

8. Discussion

We examined methods for extending inferences from randomized individuals to the target population of trial-eligible individuals, using trials nested within cohorts for outcomes observed at a single time-point post-randomization. Our work adds to the literature on "generalizability" and "transportability" (Cole and Stuart, 2010; O'Muircheartaigh and Hedges, 2014; Tipton, 2012; Hartman et al., 2013; Zhang et al., 2016; Rudolph and van der Laan, 2017), and connects with the literature on selection bias (Keiding and Louis, 2016; Infante-Rivard and Cusson, 2018).

The methods we propose rest on two identifiability conditions beyond those supported by randomization: *mean generalizability* from randomized to non-randomized individuals and *positivity of trial participation*. Directed acyclic graphs can facilitate reasoning about the mean generalizability condition (Pearl, 2015; Pearl and Bareinboim, 2014). Arguably, the identifiability conditions are most plausible in studies explicitly designed to collect information on all trial-eligible individuals, including those who are not randomized. Our methods are therefore appropriate for "comprehensive cohort studies" (Olschewski et al., 1985, 1992; Schmoor et al., 1996), randomized preference designs (Torgerson and Sibbald, 1998; Lambert and Wood, 2000), and pragmatic trials embedded within healthcare systems (Fiore and Lavori, 2016; Choudhry, 2017). Though our approach is useful for extending inferences from randomized individuals to trial-eligible individuals in these designs, it does

not address individuals who are candidates for treatment but do not meet the trial eligibility criteria.

Our approach uses treatment and outcome data only from randomized individuals, to avoid confounding of the treatment effect among non-randomized individuals and eliminate the need to follow them up. Nevertheless, if treatment and outcome data from non-randomized individuals are available, analysts can use them to assess model specification, as we did in the CASS reanalysis. When such data are available and substantive knowledge suggests that the observed covariates suffice to adjust for confounding among non-randomized individuals, it may be useful to compare the results of generalizability methods against the results of observational analyses of the entire cohort. These analyses target the same causal quantities but rest on different identifiability conditions: generalizability analyses require mean generalizability from randomized to non-randomized individuals whereas observational analyses require mean exchangeability of treated and untreated individuals. As in our CASS reanalysis, similarity of results from analyses that rest on different identifiability conditions provides mutual support for their validity (but does not definitively establish it).

Methods for generalizing inferences from randomized individuals to all trial-eligible individuals exploit models for the expectation of the outcome in the trial or the probability of trial participation. Investigators typically rely on parametric working models and model misspecification can lead to inconsistency, even if mean generalizability holds. Approximately correct specification of the model for the probability of participation may be more feasible, because inverstigators can use surveys or qualitative studies to investigate what drives eligible individuals to participate in a randomized trial (Ross et al., 1999). The findings from these investigations can be used to specify models for trial participation and obtain consistent IP weighting estimators.

Augmented IP weighting estimators are consistent when either working model is correctly specified, providing two opportunities for valid inference. Misspecification of just one of the models can adversely affect how these estimators perform. And misspecification of both models, in some cases (Waernbaum and Pazzagli, 2017), can make augmented IP weighting estimators perform worse than the (misspecified) outcome model-based estimator (Kang and Schafer, 2007). Serious bias can also occur when IP weights are highly variable, though this problem is to some extent mitigated with bounded estimators (Robins et al., 2007). Even when the outcome model is misspecified, augmented IP weighting estimators may still be preferred for their typically increased efficiency compared to non-augmented IP weighting estimators; their performance may be further improved with approaches that explicitly attempt to minimize variance (Cao et al., 2009; Rotnitzky et al., 2012) or reduce bias (Vermeulen and Vansteelandt, 2015) under model misspecification.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Bang H and Robins JM (2005). Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–973. [PubMed: 16401269]

Cao W, Tsiatis AA, and Davidian M (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723–734. [PubMed: 20161511]

Choudhry NK (2017). Randomized, controlled trials in health insurance systems. New England Journal of Medicine 377, 957–964. [PubMed: 28877014]

Cole SR and Stuart EA (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. American Journal of Epidemiology 172, 107–115. [PubMed: 20547574]

Dahabreh IJ, Hayward R, and Kent DM (2016). Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International Journal of Epidemiology 45, 2184–2193. [PubMed: 27864403]

Fiore LD and Lavori PW (2016). Integrating randomized comparative effectiveness research with patient care. New England Journal of Medicine 374, 2152–2158. [PubMed: 27248620]

Ford I and Norrie J (2016). Pragmatic trials. New England Journal of Medicine 375, 454–463. [PubMed: 27518663]

Gourieroux C, Monfort A, and Trognon A (1984). Pseudo maximum likelihood methods: Theory. Econometrica: Journal of the Econometric Society pages 681–700.

Greenland S and Robins JM (1986). Identifiability, exchangeability, and epidemiological confounding. International Journal of Epidemiology 15, 413–419. [PubMed: 3771081]

Hahn J (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–331.

Hájek J (1971). Comment on "An essay on the logical foundations of survey sampling by D. Basu". In Godambe VP and Sprott DA, editors, Foundations of Statistical Inference.

Hartman E, Grieve R, Ramsahai R, and Sekhon JS (2013). 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) 10, 1111.

Hernán MA (2004). A definition of causal effect for epidemiological research. Journal of Epidemiology and Community Health 58, 265–271. [PubMed: 15026432]

Hernán MA and Robins JM (2018). Causal inference (forthcoming). Chapman & Hall/CRC, Boca Raton, FL.

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

Infante-Rivard C and Cusson A (2018). Reflection on modern methods: selection bias?a review of recent developments. International Journal of Epidemiology.

Investigators CP (1984). 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 3, 114–128. [PubMed: 6361099]

Kang JD and Schafer JL (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science pages 523–539.

Keiding N and Louis TA (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. Journal of the Royal Statistical Society. Series A (Statistics in Society) 179, 319–376.

Lambert MF and Wood J (2000). Incorporating patient preferences into randomized trials. Journal of clinical epidemiology 53, 163–166. [PubMed: 10729688]

Little RJ and Rubin DB (2014). Statistical analysis with missing data, volume 333. John Wiley & Sons.

Lunceford JK and Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937–2960. [PubMed: 15351954]

Olschewski M, Scheurlen H, et al. (1985). Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. Methods Archive 24, 131–134.

Olschewski M, Schumacher M, and Davis KB (1992). Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. Controlled Clinical Trials 13, 226–239. [PubMed: 1320559]

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

Pearl J (2015). Generalizing experimental findings. Journal of Causal Inference 3, 259–266.

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

Robins JM (1986). 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 7, 1393–1512.

Robins JM (1999). Association, causation, and marginal structural models. Synthese 121, 151–179.

Robins JM and Greenland S (2000). Causal inference without counterfactuals: comment. Journal of the American Statistical Association 95, 431–435.

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

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

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

Ross S, Grant A, Counsell C, Gillespie W, Russell I, and Prescott R (1999). Barriers to participation in randomised controlled trials: a systematic review. Journal of Clinical Epidemiology 52, 1143–1156. [PubMed: 10580777]

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

Rotnitzky A, Lei Q, Sued M, and Robins JM (2012). Improved double-robust estimation in missing data and causal inference models. Biometrika 99, 439–456. [PubMed: 23843666]

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688.

Rudolph KE and van der Laan MJ (2017). Robust estimation of encouragement design intervention effects transported across sites. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 79, 1509–1525. [PubMed: 29375249]

Schmoor C, Olschewski M, and Schumacher M (1996). Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. Statistics in Medicine 15, 263–271. [PubMed: 8643884]

Silverman WA and Altman DG (1996). Patients' preferences and randomised trials. The Lancet 347, 171–174.

Splawa-Neyman J (1990). 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 5, 465–472.

Sun B and Tchetgen Tchetgen EJ (2018). On inverse probability weighting for non-monotone missing at random data. Journal of the American Statistical Association 113, 369–379. [PubMed: 30034062]

Tan Z (2007). Comment: Understanding OR, PS and DR. Statistical Science 22, 560–568.

Tipton E (2012). Improving generalizations from experiments using propensity score subclassification assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics 38, 239–266.

Torgerson DJ and Sibbald B (1998). Understanding controlled trials. what is a patient preference trial? BMJ: British Medical Journal 316, 360. [PubMed: 9487173]

Vermeulen K and Vansteelandt S (2015). Bias-reduced doubly robust estimation. Journal of the American Statistical Association 110, 1024–1036.

Waernbaum I and Pazzagli L (2017). Model misspecification and bias for inverse probability weighting and doubly robust estimators. arXiv preprint arXiv:1711.09388.

William J, Russell R, Nicholas T, et al. (1983). Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Circulation 68, 939–950. [PubMed: 6137292]

Williamson EJ, Forbes A, and White IR (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. Statistics in medicine 33, 721–737. [PubMed: 24114884]

Wooldridge JM (2007). Inverse probability weighted estimation for general missing data problems. Journal of Econometrics 141, 1281–1301.

Zhang Z, Nie L, Soon G, and Hu Z (2016). New methods for treatment effect calibration, with applications to non-inferiority trials. Biometrics 72, 20–29. [PubMed: 26363775]
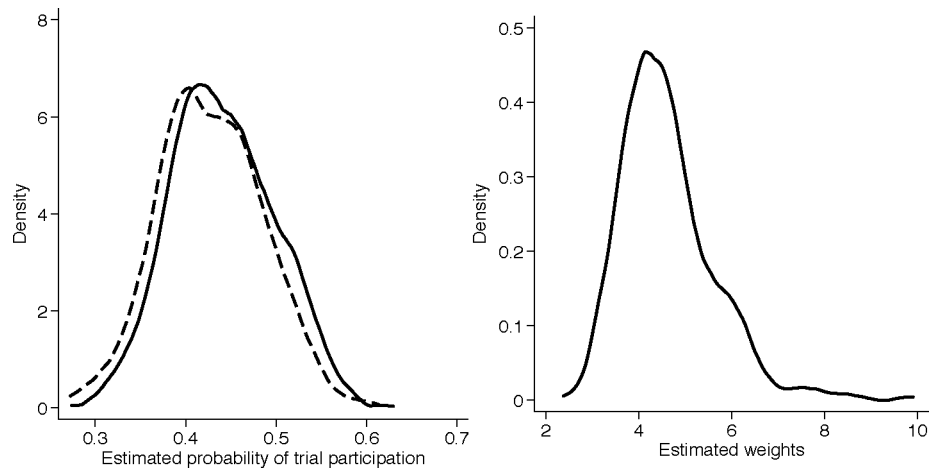
Figure 1:
Kernel densities of the estimated probabilities of trial participation (left panel) for randomized (solid line) and non-randomized (dashed line) individuals and estimated weights (right panel) for randomized individuals.

Table 1:

The left-hand-side of the table depicts the data structure, including baseline covariates X, the trial participation indicator ($S = 1$ for the $n_{S=1}$ randomized indviduals and $S = 0$ for the $n_{S=0}$ non-randomized individuals), the binary treatment (A), and the observed outcome (Y). The right hand side of the table depicts the potential outcomes Y^1 and Y^0 under treatment $a = 1$ and $a = 0$, respectively. The consistency assumption allows us to equate some of these potential outcomes with the observed outcomes, depending on trial participation and treatment assignment. Dashes denote missing values in the observed data.

Individual	X	S	A	Y	Y^1	Y^0
1	x_1	1	1	y_1	$y_1^1 = y_1$	y_1^0
2	x_2	1	1	y_2	$y_2^1 = y_2$	y_2^0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	x_{n_1}	1	1	y_{n_1}	$y_{n_1}^1 = y_{n_1}$	$y_{n_1}^0$
$n_1 + 1$	x_{n_1+1}	1	0	y_{n_1+1}	$y_{n_1+1}^1$	$y_{n_1+1}^0 = y_{n_1+1}$
$n_1 + 2$	x_{n_1+2}	1	0	y_{n_1+2}	$y_{n_1+2}^1$	$y_{n_1+2}^0 = y_{n_1+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_1 + n_0 = n_{S=1}$	$x_{n_{S=1}}$	1	0	$y_{n_{S=1}}$	$y_{n_{S=1}}^1$	$y_{n_{S=1}}^0 = y_{n_{S=1}}$
$n_{S=1} + 1$	$x_{n_{S=1}+1}$	0	–	–	$y_{n_{S=1}+1}^1$	$y_{n_{S=1}+1}$
$n_{S=1} + 2$	$x_{n_{S=1}+2}$	0	–	–	$y_{n_{S=1}+2}^1$	$y_{n_{S=1}+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_{S=1} + n_{S=0} = n$	x_n	0	–	–	y_n^1	y_n^0

Table 2:

Estimated 10-year mortality risk (%) for surgery and medical therapy, and risk difference and risk ratio comparing the treatments among randomized individuals and all trial-eligible individuals in the Coronary Artery Surgery Study. For each point estimate we provide bootstrap confidence intervals from 10,000 re-samplings. Surgery = coronary artery bypass grafting surgery plus medical therapy; Medical = medical therapy; RD = risk difference; RR = risk ratio; Trial-only = unadjusted analysis using only observations from randomized individuals; OM = outcome model-based standardization; IPW1 = inverse probability weighting; IPW2 = inverse probability weighting with normalized weights; AIPW1 = augmented inverse probability weighting; AIPW2 = augmented inverse probability weighting with normalized weights; AIPW3 = inverse probability weighted regression; (eq.) gives the equation number for each estimator.

Target population	Estimator (eq.)	Surgery (%)	95% CI	Medical (%)	95% CI	RD (%)	95% CI	RR	95% CIS
Randomized individuals	Trial-only	17.4	(13.5, 21.3)	20.1	(15.9, 24.3)	−2.7	(−8.5, 3.0)	0.86	(0.59, 1.14)
	OM (1)	18.5	(14.3, 22.6)	20.0	(15.9, 24.0)	−1.5	(−7.2, 4.2)	0.93	(0.64, 1.21)
	IPW1 (2)	18.0	(13.9, 22.1)	19.9	(15.9, 24.0)	−2.0	(−7.7, 3.8)	0.90	(0.62, 1.18)
	IPW2 (4)	18.0	(13.9, 22.0)	19.9	(15.9, 24.0)	−2.0	(−7.7, 3.8)	0.90	(0.62, 1.18)
All trial-eligible individuals	AIPW1 (5)	18.3	(14.2, 22.4)	20.0	(15.9, 24.0)	−1.6	(−7.3, 4.1)	0.92	(0.64, 1.20)
	AIPW2 (6)	18.3	(14.2, 22.4)	20.0	(15.9, 24.0)	−1.6	(−7.3, 4.1)	0.92	(0.64, 1.20)
	AIPW3 (7)	18.3	(14.2, 22.4)	19.9	(15.9, 23.9)	−1.6	(−7.2, 4.1)	0.92	(0.64, 1.20)