ANNUAL REVIEWS

Annual Review of Statistics and Its Application

A Review of Generalizability and Transportability

Irina Degtiar[1] and Sherri Rose[2]

[1]Mathematica, Inc., Cambridge, Massachusetts, USA; email: idegtiar@mathematica-mpr.com

[2]Department of Health Policy and Center for Health Policy, Stanford University, Stanford, California, USA; email: sherrirose@stanford.edu

Annu. Rev. Stat. Appl. 2023. 10:501–24

First published as a Review in Advance on October 19, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-042522-103837

Keywords

generalizability, transportability, external validity, treatment effect heterogeneity, causal inference

Abstract

When assessing causal effects, determining the target population to which the results are intended to generalize is a critical decision. Randomized and observational studies each have strengths and limitations for estimating causal effects in a target population. Estimates from randomized data may have internal validity but are often not representative of the target population. Observational data may better reflect the target population, and hence be more likely to have external validity, but are subject to potential bias due to unmeasured confounding. While much of the causal inference literature has focused on addressing internal validity bias, both internal and external validity are necessary for unbiased estimates in a target population. This article presents a framework for addressing external validity bias, including a synthesis of approaches for generalizability and transportability, and the assumptions they require, as well as tests for the heterogeneity of treatment effects and differences between study and target populations.

1. INTRODUCTION

The goal of causal inference is often to gain understanding of a particular target population based on study findings. The true underlying causal effect typically varies with the definition of the chosen target population. However, samples unrepresentative of the target population arise frequently in studies ranging from randomized controlled trials (RCTs) in clinical medicine to policy research (Bell et al. 2016). In a clinical trial setting, physicians may be left interpreting evidence from RCTs with patients who have demographics and comorbidities that are quite different from those of their patients. For example, within cancer RCTs, African Americans are widely underrepresented despite being at an increased risk for many cancers. Failing to address this lack of representation can lead to inappropriate conclusions and harm (Chen et al. 2021). In health, education, disability, or other policy settings, considering effects for the eventual target population sets expectations for anticipated results and determines groups that should be targeted for an intervention. In tech, marketing campaign A/B tests (randomized studies) that do not reflect eventual users lead to inaccurate estimates of sales revenue. Across disciplines, obtaining estimates for the target population of substantive interest, which may not align with the one in the study, better informs decision-making.

The relationships between target, study, and analysis populations are visualized in **Figure 1**. The target sample is a representative sample of the target population, whereas the study population is defined by enrollment processes and inclusion or exclusion criteria. Due to this, the study population may differ from the target population. Correspondingly, the enrolled participants who form the study sample may have different characteristics from those of the target sample. In the cancer RCT example, while a physician cares about the target population of patients that may come in to be treated at their clinic (where the clinic's current patients are a target sample), the study sample they are basing their treatment recommendations on may not include any African Americans. The study population is the hypothetical population that the study sample represents, which likewise may include no African Americans. Post-enrollment, further dropout and missingness may occur that create the observed analysis sample. In this case, patients who experienced severe adverse events may have dropped out such that the analysis sample consists of patients who did not experience severe side effects. There then exists a hypothetical analysis population from which the analysis sample data are a random sample. Hereafter, for simplicity and consistency with the literature, we use the terms study sample and study population to be inclusive of the analysis sample and analysis populations, respectively.

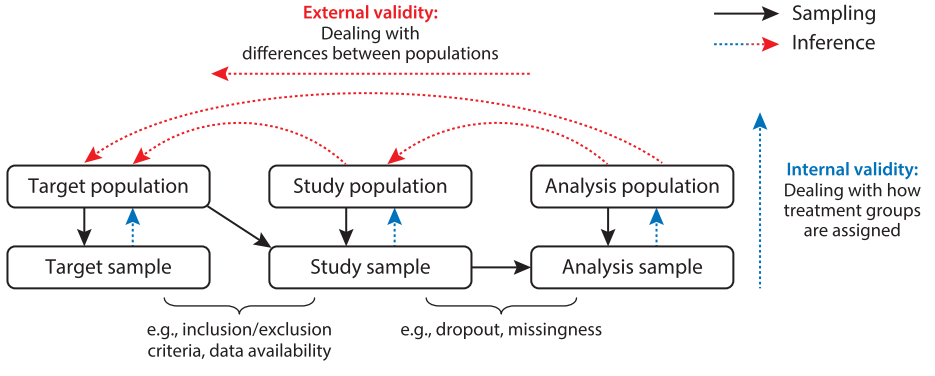

Figure 1

Internal versus external validity biases as they relate to target, study, and analysis populations.

Several key concepts underpin extending causal inferences beyond a study sample. Generalizability focuses on the setting in which the study population is a subset of the target population of interest (e.g., generalizing from a limited geography nationwide), while for transportability, the study population is (at least partly) external to the target population (e.g., transporting from one county to another). Internal validity is defined as an effect estimate being unbiased for the causal treatment effect in the population from which the sample is a simple random sample (i.e., moving vertically from a sample to its corresponding population in **Figure 1**). External validity is concerned with how well results generalize or transport to other contexts, specifically that the (internally valid) effect estimate is unbiased for the causal treatment effect in a different setting, such as a target population of interest (moving laterally between populations in **Figure 1**). External validity bias has also been referred to as sample selection bias (Heckman 1979, Imai et al. 2008, Bareinboim et al. 2014).

External validity bias arises from differences between the study and target populations in (*a*) subject characteristics; (*b*) setting, such as geography or type of health center; (*c*) treatment, such as timing, dosage, or staff training; and (*d*) outcomes, such as length of follow-up or timing of measurements (Cronbach & Shapiro 1982, Attanasio et al. 2003, Rothwell 2005). The focus of most generalizability and transportability methods is on addressing differences in subject characteristics. Hence, these methods assume the remaining threats to external validity are not present in the data sources they are looking to generalize across, e.g., in our cancer RCT example, that the clinic and teaching hospitals where the oncology trial was conducted have similar standards of cancer care and care coordination as they relate to patient outcomes, and that these outcome measures are similarly defined. Namely, external validity bias then arises solely from (*a*) variation in the probability of enrollment in the study, (*b*) heterogeneity in treatment effects, and (*c*) the correlation between items *a* and *b* (Olsen et al. 2013). We therefore distinguish between factors differentiating the target population from the study population (external validity bias) and those that create differences between treatment groups (internal validity bias), e.g., confounders. RCTs are frequently performed in a nonrepresentative subset of the target population and may have imperfect follow-up (challenging their external validity) and baseline imbalances (leading to internal validity bias). Observational studies may be susceptible to unmeasured confounding (threatening their internal validity) but may be more representative of the target population (hence having better external validity). Lack of representation in an RCT can lead to external validity bias that is larger than the internal validity bias of an observational study (Bell et al. 2016).

The optimal solution to external validity bias centers on study design, which we review only briefly here. One ideal design would randomly sample subjects from the target population and then randomly assign treatment to the selected individuals. However, this is usually infeasible. Alternative study designs for improving study generalizability and transportability include purposive sampling, in which investigators deliberately select individuals for reasons such as representation or heterogeneity (Shadish et al. 2001, Allcott & Mullainathan 2012); pragmatic or practical clinical trials, which aim to be representative of clinical practice (Schwartz & Lellouch 1967, Ford & Norrie 2016); stratified selection based on effect modifiers or propensity scores for selection (Allcott & Mullainathan 2012, Tipton 2013a); and balanced sampling designs for site selection that select representative sites through stratified ranked sampling (Tipton et al. 2017). In lieu of or in addition to study designs that address external validity bias, generalizability and transportability methods can improve the external validity of effect estimates after data collection.

This article provides a review of generalizability and transportability research, synthesizing across the statistics, epidemiology, computer science, and economics literature in a more complete manner than has been done to date. Existing review literature has examined narrower subsets of the topic: generalizing or transporting to a target population from only RCT data (Stuart et al. 2015,

Generalizability: extending causal knowledge from a study to a target population when the study population is a subset of the target population

Transportability: extending causal knowledge from a study to a target population when the study population is (at least partly) external to the target population

Confounder: factor associated with both the treatment and the outcome, which causes spurious treatment-outcome associations

Effect modifier: factor whose levels associate with different treatment effects

Propensity score: probability (of treatment assignment, study selection, etc.) conditional on measured covariates

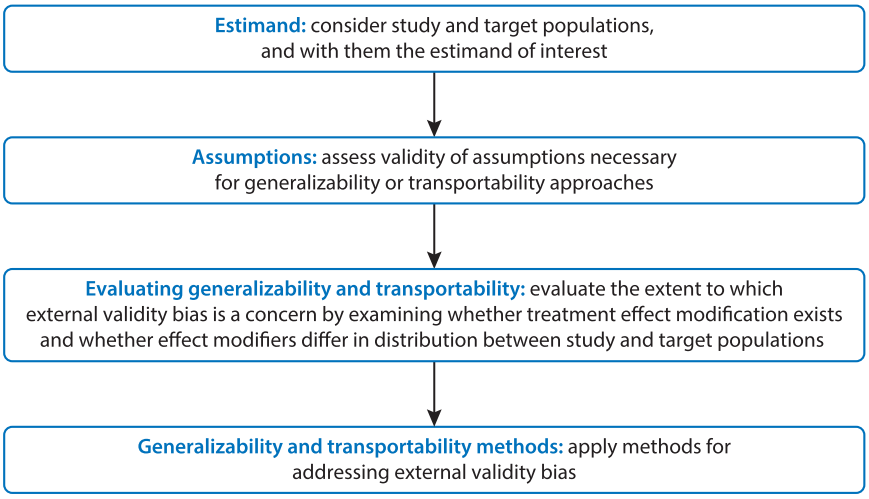

Figure 2

Framework for assessing and addressing external validity bias after data collection.

2018; Kern et al. 2016; Tipton & Olsen 2018; Ackerman et al. 2019), identifiability or concepts rather than estimation (Bareinboim & Pearl 2016; Keiding & Louis 2016, 2018), or meta-analysis approaches for combining summary-level information (Kaizar 2015, Verde & Ohmann 2015). A recent review (Colnet et al. 2021) examined combining randomized and observational data, and another (Raudenbush & Schwartz 2020) focused on approaches in education research. However, these previous reviews have not summarized the full range of generalizability and transportability methods developed across diverse disciplines that incorporate data from randomized, observational, or a combination of randomized and observational studies, nor techniques for evaluating generalizability, as we do here. Additionally, although the importance of describing generalizability and transportability is recognized by different trial reporting guidelines (e.g., CONSORT, RECORD, and STROBE), they provide no clear guidance on tests or estimation procedures (von Elm et al. 2008, Schulz et al. 2010, Benchimol et al. 2015). We also contribute recommendations for methodologists and applied researchers.

The remainder of the article synthesizes considerations for assessing and addressing external validity bias after data collection (presented as a framework in **Figure 2**) and is organized as follows. Section 2 defines the estimand of interest, the average treatment effect in a target population, as well as alternatives. Section 3 presents key assumptions underlying many of the methods. Section 4 reviews methods for assessing the extent to which external validity bias will pose a concern for the study. Namely, it reviews methods for assessing treatment effect heterogeneity, further motivating the need for methods that extend results to a target population. Section 5 then summarizes the analytic methods available for external validity bias correction that generate treatment effect estimates for a target population of interest. These techniques include weighting and matching, outcome regressions, and doubly robust approaches. Section 6 concludes with guidance for both applied and methods researchers.

2. ESTIMAND

Assume, for one or more studies, the existence of outcome Y, treatment $A \in \{0, 1\}$, and baseline covariates $\mathbf{X} \in \mathbb{R}^d$. For simplicity of notation, we define $\mathbf{X}$ to represent all potential treatment

effect confounders and effect modifiers that differ between the study and target populations. Such a simplification reflects that often, researchers assume each variable in $\mathbf{X}$ could potentially both confound and moderate effects. Within the cancer RCT setting, these could be factors such as age, previous lines of therapy, or cancer stage. Without loss of generality, we focus on the single study setting, with $S = 1$ indicating selection into it. The observational unit for the study sample is $O_{\text{study}} = (\mathbf{X}, A, Y, S = 1)$. O_{study} has probability distribution $P_{\text{study}} \in \mathcal{M}_{\text{study}}$, where $\mathcal{M}_{\text{study}}$ is our collection of possible probability distributions (i.e., statistical model). We observe n_{s} realizations of O_{study}, indexed by j. The observational unit for a representative sample from the target population is given by $O = (\mathbf{X}, A, Y, S) \sim P \in \mathcal{M}$. We have n realizations of O, indexed by i. Target sample subjects who do not appear in the study sample will have $S = 0$. In some studies, Y and A are not observed for the target sample. We use the terminology "selected" or "sampled" throughout the article for simplicity, although, for transportability, subjects are not directly sampled into the study from the target population. For generalizability, we have that $O_{\text{study}} \in O$, while for transportability, the two are disjoint sets, $O_{\text{study}} \notin O$. In our cancer RCT example, if the RCT occurred in the physician's clinic, we would be generalizing results, while if the RCT were conducted elsewhere, we would be transporting results to the clinic.

Biases are defined with respect to an estimand. We focus on the average treatment effect in a well-defined population of interest: the target population average treatment effect (PATE). Namely, we are interested in the average outcome had everyone in the target population been assigned to treatment $A = 1$ compared to if everyone had been assigned to treatment $A = 0$. We write this as $\tau = E_{\mathbf{X}}[E(Y|S = 1, A = 1, \mathbf{X}) - E(Y|S = 1, A = 0, \mathbf{X})] = E(Y^1 - Y^0)$, where Y^1 and Y^0 are the potential outcomes under treatment and no treatment, respectively, and required identifiability assumptions are as delineated in the next section. The corresponding estimator is $\hat{\tau} = 1/n \sum_{i=1}^{n} (\hat{Y}_i^1 - \hat{Y}_i^0)$. We also write Y^a to represent the potential outcome under a, with lowercase a representing a specific value for random variable A. Potential outcomes either are explicitly assumed in the potential outcomes framework or are a consequence of the structural causal model (Rubin 1974, Pearl 2000). Different target populations correspond to alternative PATEs because the expectation is taken with respect to alternative distributions of covariates $\mathbf{X}$. For example, if our clinic physician worked in both pediatrics and geriatrics, these two populations would correspond to two different PATEs. However, necessarily, we only observe outcomes in the study sample. A study therefore directly estimates the sample average treatment effect (SATE): $\tau_{\text{s}} = E(Y^1 - Y^0|S = 1)$ with estimator $\hat{\tau}_{\text{s}} = 1/n_{\text{s}} \sum_{j=1}^{n} (\hat{Y}_j^1 - \hat{Y}_j^0)$.

When the distributions of treatment effect modifiers differ between study and target populations, the true study average effect will not equal the true target population average effect (SATE $\neq$ PATE) due to external validity bias. Sampling variability as well as internal validity biases can also drive estimates of SATE further from the truth. Biases may differ in magnitude and may make the SATE either larger or smaller than the PATE. We may also be interested in estimating other target parameters. For example, some estimation methods examine the target population conditional average treatment effects (PCATEs), $\tau_x = E(Y^1 - Y^0|\mathbf{X})$, or the target population average treatment effects among the treated, $\tau_1 = E(Y^1 - Y^0|A = 1)$. Similar generalizability and transportability considerations presented in the following sections apply to these and other causal estimands.

3. ASSUMPTIONS

Under the potential outcomes framework, the assumptions below are sufficient to identify the PATE using the observed study data. A corresponding set of assumptions under the structural equation model (SEM) framework has also been derived (Pearl & Bareinboim 2011, 2014; Bareinboim & Pearl 2014, 2016; Bareinboim & Tian 2015; Pearl 2015; Correa et al. 2018).

Additional assumptions include that there are no missing data or measurement error in outcome, treatment, or covariate measurements. Other target parameters of interest necessitate a similar set of assumptions.

3.1. Internal Validity

Sufficient assumptions for identifying the PATE (and SATE) with respect to internal validity are as follows.

Assumption 1 (Conditional treatment exchangeability). $Y^a \perp A \,|\mathbf{X}, S = 1$ for all $a \in \mathcal{A}$, the set of all possible treatments.

This assumption requires no unmeasured confounding of the treatment-outcome relationship in the study. It is satisfied by perfectly randomized trials (e.g., no loss to follow-up, other informative missingness or censoring, etc.) and by observational studies that have all confounders measured. While this condition is sufficient, it is not always necessary. When estimating the PATE, it can be replaced by the weaker condition of the mean conditional exchangeability of the treatment effect, $E(Y^1 - Y^0|\mathbf{X}, A, S = 1) = E(Y^1 - Y^0|\mathbf{X}, S = 1)$ (Kern et al. 2016, Dahabreh et al. 2019b).

Assumption 2 (Positivity of treatment assignment). $P(\mathbf{X} = \mathbf{x}|S = 1) > 0 \;\Rightarrow P(A = a|\mathbf{X} = \mathbf{x}, S = 1) > 0$, with probability 1 for all $a \in \mathcal{A}$.

This assumption entails that each subject in the study has a positive probability of receiving each version of the treatment. In combination with the conditional treatment exchangeability assumption above, this assumption is also known as strongly ignorable treatment assignment (Varadhan et al. 2016).

Assumption 3 (Stable unit treatment value assumption (SUTVA) for treatment assignment). If $A = a$, then $Y = Y^a$.

This assumption requires no interference between subjects and treatment version irrelevance (i.e., consistency/well-defined interventions) in the study and target populations, respectively (Dahabreh et al. 2017, Kallus et al. 2018).

3.2. External Validity

In addition to the assumptions above, identifying the PATE involves the following parallel set of assumptions for external validity.

Assumption 4 (Conditional exchangeability for study selection). $Y^a \perp S \,|\mathbf{X}$ for all $a \in \mathcal{A}$.

This assumption is also known as exchangeability over selection and the generalizability assumption. It requires that the outcomes among individuals with the same treatment and covariate values in the study and target populations are the same (Stuart et al. 2011). All effect modifiers that differ between study and target populations must therefore be measured. This assumption would be satisfied by a study sample that is a random sample from the target population or a nonprobability study sample in which all effect modifiers are measured. In our cancer example, if cancer stage was unmeasured but modified treatment effect and the proportion of patients with stage III cancer differed between the RCT and clinic, this assumption would be violated. A weaker condition, the mean conditional exchangeability of selection, $E(Y^1 - Y^0|\mathbf{X}, S = 1) = E(Y^1 - Y^0|\mathbf{X})$,

can replace conditional exchangeability for study selection when focusing on the PATE (Kern et al. 2016, Dahabreh et al. 2019b).

Assumption 5 (Positivity of selection). $P(\mathbf{X} = \mathbf{x}) > 0 \Rightarrow P(S = 1|\mathbf{X} = \mathbf{x}) > 0$ with probability 1.

This assumption requires common support with respect to study selection; in every stratum of effect modifiers, there is a positive probability of being in the study sample or being represented by study participants (Dahabreh et al. 2017). This can be replaced by smoothing assumptions under a parametric model, for example, that the propensity score distribution for study selection has sufficient overlap or common support between the study sample and target population (Tipton & Peck 2017, Westreich et al. 2017). Thus, with conditional positivity of selection, we assume that all members of the target population are represented by individuals in the study. In our cancer RCT example, if, unlike in the clinic, there were no African American patients in the RCT, and if race was an effect modifier, this positivity assumption would be violated. However, if race was not an effect modifier, the assumption would hold. The positivity assumption in combination with the no unmeasured effect modification assumption above is also known as strongly ignorable sample selection, given the observed covariates (Chan 2017).

Assumption 6 (Stable unit treatment value assumption for study selection). If $S = s$ (and $A = a$), then $Y = Y^a$.

This assumption encompasses no interference between subjects selected into the study versus those not selected and treatment version irrelevance between study and target samples (the same treatment is given to both) (Tipton 2013b, Tipton & Peck 2017). It necessitates that there is no difference across study and target samples in how outcomes are measured or in how the intervention is applied, that there is a common data-generating function for the outcome across individuals in the study and target populations (i.e., that being in the study does not change treatment effects), and that the potential outcomes are not a function of the proportion of individuals selected for the study. For example, this assumption would be violated if, due to differential adherence, the clinic's patients received different treatment doses from those in the study or if, for an intravenous therapy, the clinic staff's training differed from that of the trial staff. Treatment version irrelevance in SUTVA can be replaced by having the same distribution of treatment versions between study and target populations when estimating the PATE (Lesko et al. 2017).

3.3. Transportability

Similar internal and external validity assumptions are needed for transportability, with the following modifications. For generalizability, the study sample is a subset of the target population; therefore, the positivity assumption for selection will need to be bounded away from 0. For transportability, the study sample is not a subset of the target population; thus, the propensity to be in the study population will need to be bounded away from 0 and 1 (Tipton 2013b). Furthermore, for transportability, the set of covariates, $\mathbf{X}$, required for conditional exchangeability for study selection cannot include those that separate the study sample from the target population (e.g., hospital type if transporting results from teaching hospitals to community clinics or geographic location if transporting between states) (Tipton 2013b). Further distinctions are discussed by Pearl (2015) using the SEM framework. Under this framework, Pearl & Bareinboim (2014) formalize the assumptions necessary for using different transport formulas to reweight randomized data, providing graphical conditions for identifiability as well as transport formulas for randomized studies (Pearl 2015), observational studies (Pearl & Bareinboim 2011, Bareinboim & Tian 2015,

Pearl 2015, Bareinboim & Pearl 2016, Correa et al. 2018), and a combination of heterogeneous studies (Bareinboim & Pearl 2014, 2016).

4. EVALUATING GENERALIZABILITY AND TRANSPORTABILITY

Numerous quantitative approaches can help evaluate the extent to which study results may be expected to extend to the target population, i.e., the extent to which external validity bias poses a concern. External validity bias exists when study and target populations differ in their distribution of effect modifiers; these assessments therefore examine population differences and whether treatment effect heterogeneity exists. Methods for assessing the similarity of study and target populations can broadly be categorized into those that compare baseline patient characteristics and those that compare outcomes for groups on the same treatment. For the former, many make use of the propensity score for selection, which also serves the purpose of assessing the extent to which propensity score adjustment using measured covariates can sufficiently remove baseline differences between study and target samples. However, most of these methods do not emphasize effect modifiers; hence, they should be combined with an assessment of whether the noted population differences correspond to heterogeneity of treatment effects. To test for heterogeneity of effects, one must first identify effect modifiers. Effect modifiers are often prespecified by the investigator, but data-driven approaches exist as well and are discussed in this section.

4.1. Assessing Dissimilarity Between Populations with Baseline Characteristics

When only summary-level study data are available, it is possible to examine differences in univariate covariate metrics between study and target samples. Cahan et al. (2017) propose a generalization score for evaluating clinical trials that incorporates baseline patient characteristics, the trial setting, protocol, and patient selection: It takes ratios of the mean or median values of these characteristics in the study and target samples and then averages across categories for an overall score. However, this approach does not account for any measures of dispersion, which may reflect the exclusion of more heterogeneous individuals from the study. When only baseline patient characteristics, and not other aspects of the study, are responsible for relevant study versus target population differences, one can perform multiplicity-adjusted univariate tests for differences in effect modifiers between study and target samples (Greenhouse et al. 2008) or examine absolute standardized mean differences (SMDs) for each covariate, $(\bar{X}_{\text{study}} - \bar{X})/\sigma_{\bar{X}}$, where $\bar{X}_{\text{study}}$ and $\bar{X}$ are the means of baseline covariates in the study and target samples, respectively, and $\sigma_{\bar{X}}$ is the standard deviation of $\bar{X}$ (Tipton & Peck 2017). High values indicate heavy extrapolation and reliance on correct model specification; in smaller samples, imbalances often occur by chance (Tipton & Peck 2017). Furthermore, these multiple comparisons can suffer from limited power. With one or more RCTs, generalizability across categorical eligibility criteria can be assessed by the percent of the target sample that would have been eligible for the study or set of studies (Weng et al. 2014); however, lack of generalizability/transportability can extend beyond eligibility.

Examining the joint (rather than the univariate) distributions of patient characteristics, such as the SMD in propensity scores for selection, more comprehensively assesses overlap (Stuart et al. 2011). When the propensity score is not symmetrically distributed, summarizing mean differences is insufficient. Tipton (2014) developed a generalizability index that bins propensity scores and is bounded between 0 and 1: $\sum_{j=1}^{k} \sqrt{w_{\text{p}_j} w_{\text{s}_j}}$, with $j = 1, \ldots, k$ bins, each with target sample proportions w_{p_j} and study sample proportions w_{s_j}. It is based on the distributions of propensity scores rather than only the averages, thus requiring patient-level study and target sample data. A generalizability index score of <0.5 suggests a study being very challenging to generalize from, and a score of >0.9 suggests high generalizability (Tipton 2014). Other propensity score distance

measures can be used, such as Q-Q plots and the Kolmogorov-Smirnov distance, Levy distance, overlapping coefficient, and C statistic; these largely focus on comparing cumulative densities (Tipton 2014, Ding et al. 2016). To assess the degree of extrapolation, one can examine overlap in the propensity of selection distributions, such as the proportion of target sample individuals with propensity scores outside the 5th and 95th percentiles of the sample propensity scores (Tipton & Peck 2017). For example, Tipton (2014) uses the generalizability index to demonstrate that an educational intervention could be reasonably generalized to the United States as a whole as well as three states individually, but generalization to other states may be problematic.

The machine learning literature provides an alternative approach: detecting covariate shift—a change in the distribution of covariates between training and test data (here, the study and target data) (Glauner et al. 2017). After creating a joint data set with target and study sample data, a classification algorithm predicts whether the data came from the study. A dissimilarity metric surpassing a threshold of acceptability then suggests sizable dissimilarity between data sets. However, an inability to accurately predict study versus target data origin does not rule out differences in effect modifiers. A low score might furthermore indicate an incorrect model specification or insufficient model tuning.

These tests assess differences between populations; however, they require knowledge of which characteristics moderate the treatment effect (or are correlated with unmeasured effect modifiers) and what level of differences are substantively relevant. Many covariates are often tested or included in a propensity score regression for study selection. This approach prioritizes predictors that are strongly associated with study selection rather than those that exhibit strong effect modification. Investigators should aim to identify relevant effect modifiers for testing or inclusion in the propensity score regression and test this subset.

4.2. Assessing Dissimilarity Between Populations with Outcomes

When individual-level outcome data or joint distributions of group-level outcome data are available in both the study and target samples for at least one of the treatment groups, the following methods can assess the extent to which measured effect modifiers account for population differences. One can compare the observed outcomes in the target sample to predicted outcomes using study controls (Stuart et al. 2011) or, more generally, study individuals who received the same treatment (Hotz et al. 2005): $1/n_a \sum_{i=1}^{N} 1(A_i = a)Y_i$ versus $1/n_{s,a} \sum_{i:S_i=1} 1(A_i = a)w_i Y_i$, with weights w_i defined by weighting and matching methods discussed in Section 5.1. Hartman et al. (2015) formalize this comparison with equivalence tests. Alternatively, conditional outcomes for study and nonstudy target sample individuals receiving the same treatment, conditioning on measured effect modifiers, can be compared to detect unmeasured effect modification: $\hat{E}(Y|\mathbf{X}, A = a, S = 1)$ versus $\hat{E}(Y|\mathbf{X}, A = a, S = 0)$, although other identifiability assumption violations might also be at fault. Possible tests include analysis of covariance, Mantel-Haenszel, U-statistic-based tests, stratified log-rank, or stratified rank-sum, depending on the outcome (Marcus 1997, Hotz et al. 2005, Luedtke et al. 2019). For example, study controls could be compared to subgroups of the target population that were known to be excluded from the study (e.g., patients who declined participation in our cancer RCT example). Relatedly, unmeasured effect modification can be imperfectly tested for by disaggregating a characteristic that differentiates the study from the target sample (Allcott & Mullainathan 2012). These outcome differences should not exceed those observed between study treatment groups (Begg 1992).

Rather than testing for outcome differences, to assess the sufficiency of the estimation approach, one can test for differences between study and target regression coefficients or between baseline hazards in a Cox regression (Pan & Schaubel 2009). Any identified differences in

outcomes or effects will reflect sample differences unaccounted for by the outcome or weighting method, indicating unmeasured effect modification or an ineffective modeling approach. To have this comparison reflect relevant differences, study controls must be representative of the target population after weighting or regression adjustment. Hartman et al. (2015) provide a more formal set of identifiability assumptions that may be violated when each equivalence test is rejected. If unmeasured effect modification is suspected, one can perform sensitivity analysis to assess the extent to which it can impact results (Marcus 1997, Nguyen et al. 2017, Dahabreh et al. 2019c) or to generate bounds on the treatment effect when only partial identification is possible (Chan 2017).

4.3. Identifying Treatment Effect Heterogeneity

Identified population differences are relevant insofar as they correspond to differences in treatment effect modifiers. The following tests enable an investigator to assess whether treatment effects vary substantially across measured covariates. Many are suitable for use in observational or RCT data, although they have largely been demonstrated in RCT data to date. While some tests require a priori specification of subgroups, others can discover them in data-driven ways and most require individual-level data. A straightforward but often overlooked issue is that for studies whose participants are homogeneous with respect to effect modifiers, investigators will have difficulty identifying heterogeneity of effects. These approaches are therefore best applied to data representative of the target populations (Gunter et al. 2011).

Tests of prespecified subgroups should focus on target population subgroups under- or overrepresented in the study, or any other substantively relevant subgroup expected to exhibit effect heterogeneity. Methods for testing treatment effect heterogeneity of a priori specified subgroups largely exhibit limited power. Those testing several effect modifiers individually are particularly underpowered to detect significant effects after incorporating multiple testing adjustments, e.g., testing the interaction terms of treatment assignment with effect modifiers in a linear model, which also requires modeling assumptions regarding the linearity and additivity of effects (Gabler et al. 2009, Fang 2017). To address this lack of power, sequential tests for identifying treatment-covariate interactions can be used with either randomized or observational data (Qian et al. 2019). Alternative approaches, each addressing slightly different goals, include testing whether the conditional average treatment effect is identical across predefined subgroups (Crump et al. 2008, Green & Kern 2012), comparing subgroup effects to average effects (Simon 1982), and identifying qualitative interactions or treatment differences exceeding a prespecified relevant threshold (Gail & Simon 1985).

When effect modifiers are not known a priori, a variety of techniques identify subgroups with heterogeneous effects. These include those that identify variables that qualitatively interact with treatment (i.e., for which the optimal treatment differs by subgroup) (Gunter et al. 2011) as well as determine the magnitude of interaction (Tian et al. 2014, Chen et al. 2017). Various machine learning approaches also identify subgroups with heterogeneous treatment effects while minimizing modeling assumptions. Approaches that also present tests for treatment effect differences between subgroups include Bayesian additive regression trees (BARTs) and other classification and regression tree (CART) variants (Su et al. 2008, 2009; Green & Kern 2012; Athey & Imbens 2016). Tree-based methods develop partitions in the covariate space recursively to grow toward terminal nodes with homogeneity for the outcome. These approaches may be particularly useful when heterogeneity may be a function of a more complex combination of factors. This could be the case in our cancer RCT example if we consider social determinants of health.

With many effect modifiers, or when effect modifiers are unknown, global tests for heterogeneity can be used. Pearl (2015) provides conditions for identifying treatment effect heterogeneity

(including heterogeneity due to unmeasured effect modifiers) for randomized trials with binary treatments, situations with no unobserved confounders, and studies using mediating instruments. Effect heterogeneity can be tested for using the baseline risk of the outcome as an effect modifier; interaction-based tests assess for differences in baseline risk between study and target control groups (Weiss et al. 2012, Varadhan et al. 2016). These tests avoid multiple testing but require target sample outcome data and modeling assumptions. A consistent nonparametric test also exists that assesses for constant conditional average treatment effects, $\tau_{\mathbf{x}} = \tau \; \forall \mathbf{x} \in \mathcal{X}$ (Crump et al. 2008). Additional methods, which suffer from limited power and rely on estimates of SATEs, include testing whether potential outcomes across treatment groups have equal variances and whether cumulative distribution functions of treatment and control outcomes differ by a constant shift (Fang 2017). Global tests do not identify effect modifiers, although if a global test is rejected, one can then compare individual subgroups to determine which demonstrate effect heterogeneity.

If these assessments of generalizability fail and the target population is not well-represented by the study population (specifically, when strong ignorability fails), Tipton (2013b) provides several recommended paths forward. Investigators can change the target population to one represented by the study, that is, change the estimand of interest by aligning inclusion and exclusion criteria, outcome time points, or treatment doses (Hernán et al. 2008, Weisberg et al. 2009). A population coverage percentage can then summarize the percent overlap between the new and original target sample propensity scores and describe relevant differences from the original target population. Investigators can alternatively retain the original target population and note the limitations of extrapolated results and likelihood of remnant bias. However, a different study may need to be conducted instead.

5. METHODS FOR ESTIMATING POPULATION AVERAGE TREATMENT EFFECTS

Following the application of the methods in the previous sections, including assessing the plausibility of relevant assumptions, an analytic method is typically needed to generalize or transport results from randomized or observational data to a target population. These approaches have many parallels to those used to address internal validity bias. We revisit matching- and weighting-based methods and outcome regressions in depth while additionally examining techniques that use both propensity and outcome regressions (these are often doubly robust). To mitigate external validity bias, generalizability and transportability methods address differences in the distribution of effect modifiers between study and target populations. To do so, for matching- and weighting-based approaches, these methods account for the probability of selection into the study rather than the probability of treatment assignment. Outcome regressions require that the treatment effect is allowed to vary across all effect modifiers in addition to all confounders being correctly included in the regression.

Most generalizability and transportability methods have been developed for randomized data. When outcome data are available from both randomized studies and an observational study representative of the target population, their combination has the potential to overcome sensitivity to positivity violations for selection into the study (an issue that RCT data commonly face) as well as to unmeasured confounding (which may afflict observational studies). Incorporating observational data in a principled manner can also shrink the mean squared error. However, many such approaches do not leverage the internal validity of RCT data. The following sections highlight some exceptions. While most approaches require individual-level study and target sample data, we also highlight approaches that use only summary-level data for either the study or the target sample.

5.1. Matching and Weighting Methods

Methods that adjust for differing baseline covariate distributions between study and target samples via matching or weighting are particularly effective when effect modifiers strongly predict selection into the study. While including unnecessary covariates can decrease precision, increase the chance of extreme weights and difficult-to-match subjects, and provide no bias reduction (Nie et al. 2013), failing to include an effect modifier is typically of greater concern than including unnecessary covariates (Stuart 2010, Dahabreh et al. 2020). Matching and reweighting methods strongly rely on common covariate support between study and target populations and perform poorly when a portion of the target population is not well-represented in the study sample or when empirical positivity violations occur. Investigators should use the estimation approach that leads to the best effect modifier balance for their study (Stuart 2010) and strive for fewer assumptions.

5.1.1. Matching. Full matching and fine balance of covariate means have been used in the generalizability context (Stuart et al. 2011, Bennett et al. 2020). Stuart et al. (2011) fully match study and target sample individuals based on propensity scores to form sets, each with at least one study and target individual. Individuals' outcomes are then reweighted by the number of target individuals in their matched set. This approach relies heavily on the distance metric, which can be misled by covariates that do not affect the outcome. Matching with fine balance is a computationally efficient nonparametric approach that accommodates multivalued treatments (Bennett et al. 2020). This approach matches samples to a target population to achieve fine balance on covariate means rather than working with the propensity score. For our cancer therapy example, each clinic patient would be matched to a treated and a control patient in the RCT to attain balance on the marginal distributions of **X**s. Some implementations of these methods only match a subset of study individuals (and hence show areas of the covariate distribution without common support), while others ensure all study and target sample individuals are matched. Matching methods require calibration for bias-variance trade-off such as via a caliper or by choosing the ratio of study to target individuals to match. A variety of distance metrics exist; however, none specifically target effect modifiers.

5.1.2. Weighting. In a low-dimensional setting with categorical or binary covariates, nonparametric poststratification (also known as direct adjustment or subclassification) has been used with randomized data (Miettinen 1972, Prentice et al. 2005) and with observational data in the context of instrumental variables (Angrist & Fernández-Val 2013). Poststratification obtains estimates for each stratum of effect modifiers, then reweights these estimates to the effect modifier distribution in the target population, i.e., $\hat{E}(Y^a) = 1/n \sum_{l=1}^{L} n_l \bar{Y}_l^a$, where L is the number of strata; n_l is the target sample size in stratum l, $n = \sum_{l=1}^{L} n_l$; and $\bar{Y}_l^a$ is an estimate from study sample data of the potential outcome on treatment a in stratum l, commonly the stratum-specific sample mean for subjects on treatment a (Miettinen 1972, Prentice et al. 2005). For example, one would obtain separate cancer therapy impact estimates for each combination of cancer stage, age category, and previous lines of therapy, then take a weighted average, with weights corresponding to the proportion of clinic patients falling into each stratum. Poststratification only requires stratum-specific summary data, and closed-form variance formulas are often available. However, empty strata are an issue with continuous variables or many stratifying variables. Conversely, if insufficient strata are used, residual external validity bias remains, which is particularly problematic in small samples (Tipton & Peck 2017). To combat this, inference can be pooled across strata using multilevel regression with poststratification (Pool et al. 1964, Gelman & Little 1997). For higher-dimensional settings or with continuous covariates, flexible nonparametric approaches can be applied, such as maximum entropy weighting, which reweights study data to the target sample distribution

(Hartman et al. 2015). When target and study populations differ on posttreatment variables such as adherence, principal stratification can be used by classifying subjects into never-taker, always-taker, and complier categories (Frangakis 2009).

Most weighting approaches use a propensity of study selection regression to construct weights. They rely on correct specification of the propensity score regression and sufficient propensity score overlap between study subjects and target sample individuals not in the study. These approaches have the advantage of allowing one set of weights to be used for treatment effects related to multiple outcomes. The most straightforward weighting approaches tend to have large variances in the presence of extreme weights, give disproportionate weight to outlier observations, and produce outcome estimates outside the support of the outcome variable. Weight standardization can address these issues, as can weight trimming, although the latter induces bias by changing the target population of interest, hence requiring a careful bias-variance trade-off.

Inverse probability of participation weighting (IPPW), a Horvitz-Thompson-like approach (Horvitz & Thompson 1952), is the most common weighting technique for generalizability (Flores & Mitnik 2013; Lesko et al. 2017; Westreich et al. 2017; Correa et al. 2018; Dahabreh et al. 2019b, 2020). Most simply, IPPW weights the outcome for each study individual on treatment a by the inverse probability (propensity) of being in the study. Weights have been developed for estimating PATEs, including those that incorporate treatment assignment to account for covariate imbalances in an RCT or for confounding in an observational study. The observed outcomes are reweighted to obtain the potential outcomes for each treatment group a: $E(Y^a) = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i$, with

$$w_i = \frac{1}{\pi_{s,i}} I(S_i = 1) I(A_i = a)$$

for random treatment assignment (Lesko et al. 2017) and

$$w_i = \frac{1}{\pi_{s,i}\pi_{a,i}} I(S_i = 1) I(A_i = a)$$

more generally (Stuart et al. 2011, Dahabreh et al. 2019b). Here $I(S_i = 1)$ is the indicator for being in the study, $I(A_i = a)$ is for being assigned treatment a, $\pi_{s,i} = P(S_i = 1|X_i)$ is the propensity score for selection into the study, and $\pi_{a,i} = P(A_i = a|S_i = 1, X_i)$ is the propensity score for assignment to treatment a in the study.

Individual-level data are typically required, although one can also use joint covariate distributions from group-level data (Cole & Stuart 2010) or univariate moments (e.g., means and variances) with additional assumptions (Signorovitch et al. 2010, Phillippo et al. 2018). For example, Cole & Stuart (2010) use the cross-classification of sex, race, and age groups in a human immunodeficiency virus trial and in the US population to fit a propensity for selection into the trial. Because IPPW only uses study individuals on a given treatment to estimate potential outcomes for that treatment, power can become an issue, particularly for multilevel treatments. These methods also perform poorly when study selection probabilities are small, a common occurrence for generalizability (Tipton 2013b). IPPW has also been developed for regression parameters in a generalized linear model (Haneuse et al. 2009).

For transportability to the target population $S = 0$, odds of participation weights are used rather than inverse probability of participation weights (Westreich et al. 2017, Dahabreh et al. 2020). This corresponds to the estimator $E(Y^a|S = 0) = \frac{1}{n}\sum_{i=1}^{N} w_i Y_i$ with $N = n + n_s$ and weights $w_i = \frac{1-\pi_{s,i}}{\pi_{s,i}\pi_{a,i}} I(S_i = 1) I(A_i = a)$ (Dahabreh et al. 2020). To address potentially unbounded outcome estimates, standardization then replaces n by the sum of the weights, which normalizes the weights to sum to 1 (Dahabreh et al. 2019b, 2020). The resulting estimator is more stable, is bounded by

the range of the observed outcomes, and performs better when the target sample is much larger than the study.

Under regularity conditions, estimates derived using IPPW are consistent and asymptotically normal (Lunceford & Davidian 2004, Cole & Stuart 2010, Buchanan et al. 2018, Correa et al. 2018). Variance can be obtained through either a bootstrap approach or robust sandwich estimators. The latter may be difficult to calculate (Haneuse et al. 2009), and bootstrap methods for IPPW have been shown to perform better when there is substantial treatment effect heterogeneity or smaller sample sizes (Tipton & Peck 2017).

Propensity scores can also be used in the context of poststratification, weighting or matching individuals within strata. RCT individuals are divided into strata defined by their propensity scores; quintiles are commonly used, based on results showing that this approach may remove over 90% of bias (O'Muircheartaigh & Hedges 2014). Effects are estimated using sample data within each subgroup, such as through separate regressions or a joint parametric regression with fixed effects for subgroups and interaction terms for subgroups by RCT status. Results can then be reweighted based on the number of target sample individuals in each subgroup (O'Muircheartaigh & Hedges 2014). Alternatively, the target sample can be matched to RCT individuals within the same propensity score stratum (Tipton 2013b).

The poststratification estimator is asymptotically normal and closed-form variance estimates exist for independent strata (Lunceford & Davidian 2004, O'Muircheartaigh & Hedges 2014). Compared to IPPW, strata reweighting is more likely to be numerically stable and easily implementable when treatment assignment is done at the group level (e.g., cluster-randomized trials). However, stratification implicitly assumes that treatment effects are identical for study and target patients in the same stratum; this assumption is rarely met, resulting in residual confounding and inconsistent estimates (Lunceford & Davidian 2004). It also relies on the assumptions that treatment effect heterogeneity is fully captured by the propensity score for treatment and that outcomes are continuous and bounded. With too few strata, bias reduction will be insufficient; conversely, too many strata can lead to small strata counts and unstable estimates (Stuart 2010, Tipton & Peck 2017).

5.2. Outcome Regression Methods

Outcome regressions, also known as response surface modeling, have not been as extensively developed for generalizability and transportability compared to propensity-based approaches. We highlight approaches that combine outcome data from one or multiple studies.

5.2.1. Outcome data from one study. Outcome regression approaches fit an outcome regression in study sample data to estimate conditional means, then obtain PATEs by marginalizing over (i.e., standardizing to) the target sample covariate distribution via predicting counterfactuals for the target sample: $\hat{E}(Y^a) = \frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y_i|S_i = 1, A_i = a, X_i)$. If the target sample is not a simple random sample from the target population, this would be a weighted average using sampling weights (Kim et al. 2018).

Outcome regression approaches are particularly effective when effect modifiers strongly predict the outcome and when the outcome is common but selection into the study is rare. They are also convenient for exploring PCATEs. These methods can yield better precision than weighting- or matching-based methods because they can adjust for confounders, effect modifiers, and factors only predictive of the outcome, thus decreasing variance in the estimate. They are simple to implement when an outcome regression for confounding adjustment has already been fit and accounts for all relevant effect modifiers. The same regression that was used to estimate impacts within the study can then be used to predict counterfactuals in the target sample. Outcome regression

methods can be used with either randomized or observational study data but have been used most frequently in RCTs. In the presence of significant lack of overlap between the target and study samples, outcome regressions rely on heavy extrapolation (Attanasio et al. 2003, Kern et al. 2016), often with no corresponding inflation of the variance to reflect uncertainty in the resulting estimates.

The simplest approach is ordinary least squares regression (Flores & Mitnik 2013; Kern et al. 2016; Dahabreh et al. 2019b, 2020). An outcome regression is fit with interaction terms between treatment and all effect modifiers before predicting counterfactual outcomes for the target sample. Dahabreh et al. (2020) show the consistency of this type of outcome regression for the PATE. For RCTs, separate regressions are recommended for each treatment group to better capture treatment effect heterogeneity (Dahabreh et al. 2019b), although this approach precludes borrowing information across treatment groups, which is possible with machine learning methods that discover treatment effect heterogeneity. Among these machine learning techniques is BART, which is the most commonly used data-adaptive outcome regression approach for generalizability and transportability (Chipman et al. 2007, 2010; Hill 2011; Kern et al. 2016). BART models the outcome as a sum of trees with linear additive terms and a regularization prior. It addresses external validity bias via its data-driven discovery of treatment effect heterogeneity; strengths of the method include its ability to obtain confidence intervals from the posterior distribution (Hill 2011, Green & Kern 2012). However, BART credible intervals show undercoverage when the target population differs substantially from the RCT (Hill 2011). Data availability may challenge these outcome regression approaches. When the covariates in the target sample are not available in the study sample, or vice versa, but the SATE is expected to be approximately unbiased for the PATE, the SATE estimates' credible intervals can be expanded to account for uncertainty in the target population covariate distribution (Hill 2011).

5.2.2. Outcome data from multiple studies. Here, we consider meta-analytic approaches for summary-level data as well as studies that combine individual-level data from more than one study (for example, one randomized and one observational study). Much of the literature has focused on meta-analytic techniques using summary-level study data and no target sample covariate information. This body of bias-adjusted meta-analysis methods largely does not explicitly define a target population for whom inference is desired but rather relies on subjective investigator judgments of each study's levels of bias, specified using bias functions or priors in a Bayesian framework. Eddy (1989) presents the first such approach, which adjusts each study for (investigator-specified) internal and external validity biases. Subsequent Bayesian hierarchical models include Prevost et al.'s (2000) three-level model, which addresses variability between studies, study types (randomized versus observational), and effect heterogeneity but does not explicitly consider internal and external validity biases (Kaizar 2011).

Other meta-analysis studies leveraging summary-level data separately specify and subjectively quantify internal and external validity bias parameters for an explicit target population, downweighting studies with higher risk of bias [e.g., Turner et al.'s (2009) bias-adjusted meta-analysis checklist approach or Greenland's (2005) Bayesian meta-sensitivity model with bias parameters for misclassification, nonresponse, and unmeasured confounding]. In the intermediate setting in which individual-level data are available in the study but only covariate moments (e.g., means, variances) are available in the target setting, Phillippo et al. (2018) present an outcome regression approach for indirect treatment comparison across RCTs.

When individual-level outcome data are available in the target sample or from multiple studies, data can be combined into one joint data set for outcome regression analysis (Kern et al. 2016). Such an approach can be preferential to IPPW, which uses only study and not target sample outcome data (Kern et al. 2016). However, it will be dominated by observational data results (and

their potential biases) when observational subjects constitute the majority of the joint data set, effectively resulting in a weighted average across studies.

Hierarchical Bayesian evidence synthesis is the only outcome regression approach we identified that attempts to empirically adjust for unobserved confounding when estimating effects for observational patients who are not well-represented in the RCTs (Verde et al. 2016). Summary-level RCT data are combined with individual-level observational data through a weighting approach in which the control group event rate is assumed to be similar across all studies and a study quality bias term is added to the observational studies' outcome regression to account for unmeasured confounding or other uncontrolled biases and to inflate variance. Verde et al. (2016) apply this approach to extrapolate results from six RCTs to a cohort study of a diabetic foot ulcer therapy. Alternatively, Gechter (2015) derives bounds on the PATE and PCATE when transporting from an RCT to a target sample with outcome data (all untreated).

5.3. Combined Propensity Score and Outcome Regression Methods

Double robust methods for generalizability and transportability typically combine outcome and propensity of selection regressions. They are asymptotically unbiased when at least one of these regression functions is consistently estimated, and if both are consistently estimated, asymptotically efficient. Incorporating flexible modeling approaches can help mitigate regression misspecification.

5.3.1. Outcome data from one study. Three asymptotically locally efficient double robust approaches have been developed in randomized data: a targeted maximum likelihood estimator (TMLE) for instrumental variables (Rudolph & van Der Laan 2017), which is a semiparametric substitution estimator, the estimating equation-based augmented inverse probability of participation weighting (A-IPPW) (Dahabreh et al. 2019b, 2020), and an augmented calibration weighting estimator that can also incorporate outcome information from the target sample when it is available (Dong et al. 2020).

The TMLE was developed for transportability in an encouragement design setting (i.e., intervention focused on encouraging individuals in the treatment group to participate in the intervention) with instrumental variables (Rudolph & van Der Laan 2017) and has also been used for generalizability (Schmid et al. 2020). Three different PATE estimators were developed: intent to treat, complier, and as treated. All use an outcome regression to obtain an initial estimate then adjust that estimate with a fluctuation function using a clever covariate C, which is derived from the efficient influence curve and incorporates the propensity of selection information in a bias-reduction step. For example, for the intent to treat PATE, the fluctuation function takes the form $\text{logit}(\hat{E}(Y|S=1,A,Z,\mathbf{X})+\epsilon C)$, where

$$C=\frac{I(S=1,A=a)}{P(A=a|S=1,\mathbf{X})P(S=1)}\frac{P(Z=z|S=0,A=a,\mathbf{X})P(\mathbf{X}|S=0)}{P(Z=z|S=1,A=a,\mathbf{X})P(\mathbf{X}|S=1)},$$

Z is the intervention taken, and A is the assigned intervention.

A-IPPW has been developed for generalizing results to estimate PATEs for all trial-eligible individuals (Dahabreh et al. 2019a,b) and transporting results to estimate PATEs for trial-eligible individuals not included in a trial (Dahabreh et al. 2020). Three estimating equation-based estimators are presented: A-IPPW, A-IPPW with normalized weights to ensure bounded estimates, and a weighted outcome regression estimator using participation weights. With w_i defined as before, the nonnormalized A-IPPW estimators are

$$\frac{1}{n}\sum_{i=1}^{n}\{w_i\{Y_i-\hat{E}(Y_i|S_i=1,A_i=a,X_i)\}+\hat{E}(Y_i|S_i=1,A_i=a,X_i)\}$$

for generalizability and

$$\frac{1}{n}\sum_{i=1}^{N}\{w_i\{Y_i - \hat{E}(Y_i|S_i = 1, A_i = a, X_i)\} + \{1 - I(S_i = 1)\}\hat{E}(Y_i|S_i = 1, A_i = a, X_i)\}$$

for transportability. Variance can be derived using empirical sandwich estimates or using a nonparametric bootstrap. As these estimators are partial M-estimators, they can produce estimates outside bounds if the outcome regression is not well chosen, and they may have multiple solutions.

Several other double robust estimators for transportability resemble the IPPW estimator, with sampling weights derived through alternative approaches that do not rely on propensity scores (Dong et al. 2020, Josey et al. 2021). For example, the semiparametric and efficient augmented weighting estimator by Dong et al. (2020) calibrates the RCT covariate distribution to match that of the sampling-weighted target sample.

An alternative reweighted outcome regression method for observational data does not claim double robustness; this regularized neural network estimator for PCATE parameters jointly learns representations from the data and a reweighting function (Johansson et al. 2018). Representational learning creates balance between the study and target covariate distributions and between treated and control distributions in a representational space so that predictors use information common across these distributions and focus on covariates predictive of the outcome. In this learned representational space, results are then reweighted to minimize an upper bound on the expected value of the loss function under the target covariate distribution.

5.3.2. Outcome data from multiple studies. Several methods have combined randomized and observational data sources such that they retain internal and external validity. These approaches broadly rely on the assumption that the relationship between unmeasured confounders and potential outcomes is the same in the RCT as in the target sample, which is a weaker assumption than that of no unmeasured confounding required by most of the methods described thus far. One study combined individual-level data from several RCTs to transport results to the target sample, extending the A-IPPW estimator (as well as corresponding IPPW and outcome regression estimators) to the multistudy setting (Dahabreh et al. 2022). The remaining estimators in this section combine randomized and observational data.

When differences in effect modifiers between the RCT and target population are known (e.g., by inclusion and exclusion criteria), cross-design synthesis meta-analysis is a method for combining randomized and observational study data while capitalizing on the internal validity of the randomized data and the external validity of the observational data (Begg 1992, Greenhouse et al. 2017). It provides a means for estimating treatment effects for patients excluded from the RCT and can use summary-level RCT data if outcomes are available by relevant patient subgroups, although it can only accommodate a limited number of strata of relevant effect modifiers. Cross-design synthesis meta-analysis assumes a constant amount of unmeasured confounding across patients eligible and ineligible for the RCTs (Kaizar 2011).

When differences between RCT and target populations are less well understood, there are continuous effect modifiers, or a higher-dimensional set of effect modifiers exists, one can use Bayesian calibrated risk-adjusted regressions (Varadhan et al. 2016). This parametric approach requires individual-level information from observational and randomized studies, leveraging outcome regressions and calibration using the propensity of selection. The target population is assumed to be represented by a subset of the observational data; the RCT data are likewise assumed to be represented by a (potentially different) subset of the observational data. The method relies on the observational data set having substantial effect modifier overlap with both the target sample and the RCT.

A two-step frequentist approach for consistently estimating PCATE parameters has also been developed (Kallus et al. 2018). It begins with outcome regressions for each treatment group of the observational data, or a flexible regression that captures effect heterogeneity. Observational data are then standardized to the RCT population before so-called debiasing their estimates using RCT data by including a correction term that can depend on measured covariates. This method relies on the assumption that calibrating internal validity bias in the subset of the observational data distribution overlapping with RCT data appropriately calibrates the bias for the entire target sample. The approach would therefore not necessarily decrease bias if the covariate distribution is highly imbalanced, resulting in average biases that are different between the RCT overlapping versus nonoverlapping subsets of the target sample. Degtiar et al. (2021) overcome this limitation by standardizing to the observational data itself when debiasing observational data estimates. The approach accommodates outcome regression, propensity weighting, and double robust estimators. Lu et al. (2019) present a semiparametric double robust approach that, unlike the above methods, assumes no unmeasured confounding in the observational data when combining RCT and comprehensive cohort study data (in which patients who decline randomization are enrolled in a parallel observational study).

6. DISCUSSION

Obtaining unbiased estimates for a relevant target population requires applying generalizability or transportability methods in studies that meet required identifiability assumptions. The internal validity of randomized trials is not sufficient to obtain unbiased causal effects—external validity also needs to be considered. In this synthesis, we have discussed (*a*) sources of external validity bias and study designs to address it; (*b*) the definition of an estimand in a target population of interest; (*c*) the identifiability assumptions underpinning generalizability and transportability approaches; (*d*) a variety of approaches for quantifying the relevant dissimilarity between study and target samples and assessing treatment effect heterogeneity; and (*e*) a variety of matching and weighting methods, outcome regression approaches, and techniques that use both outcome and propensity regressions that generalize or transport from randomized and observational studies to a target population. These approaches have been applied across diverse settings from RCT results transported to patients represented in registries to cluster-randomized educational intervention trials generalized to broader geographic areas. Across a variety of settings, it is important to estimate results for populations that go beyond the study population, and we suggest the following considerations.

- Make efforts to explicitly define target population(s) and identify the study population from which the study sample is a random sample. While describing the study population may be difficult, and there may not be a practically meaningful population representative of the study sample data, this clarity will allow one to compare and, when feasible, better align the study sample to the target population. Discussion regarding target population(s) should be guided by ensuing decisions the study aims to inform as well as practical considerations (e.g., lack of certain subgroups in the study). These considerations may require iteration between feasibility and desired study aims. When combining studies, meta-analyses should likewise carefully specify target population(s) for inference and incorporate considerations of treatment effect heterogeneity or demonstrate that it is not a concern. Without transparency in the target population(s), a study cannot estimate well-defined treatment effects, nor can readers judge the generalizability of study results to any other population of interest.

- Plan for generalization in the study design, when feasible including writing generalizability considerations into grant or study objectives. Enroll randomized study participants or design observational study inclusion and exclusion criteria to have the study sample be representative of the target population or fully capture the heterogeneity of effect modifiers. Collect data on likely treatment effect modifiers that are associated with study participation. Attempt to identify and mitigate potential sources of missingness or selection bias. If possible, collect baseline characteristics and outcome data on study nonparticipants who are part of the target population. Otherwise, identify external sources of data that might inform the composition of the target population with respect to effect modifiers and work toward aligning variables between these target sample data sources and the study.
- Clearly describe the internal and external validity assumptions needed to identify the treatment effect as they relate to the study. Substantively assess the justifiability of these internal and external validity assumptions. To the extent possible, test the validity of the assumptions and perform sensitivity analyses to assess the impact of assumption violations.
- Quantify the dissimilarity between the study and target populations using at least one method. Ideally, use multiple methods, as they each tell different parts of the story: Examine univariate and joint distributions of effect modifiers, differences in the propensity to participate in the study, and (if outcome information is available in the target sample) differences in outcomes between study and target subjects on the same treatment. If differences are identified, one should investigate which subpopulations drive those differences and assess whether they have heterogeneous treatment effects. In addition to examining subject characteristics, assess whether differences exist in the setting, treatment, or outcome measure.
- To obtain causal estimates when the target and study populations differ with respect to effect modifiers, incorporate at least one generalizability or transportability estimator. Alternatively, at the minimum, assess and describe sources of effect heterogeneity and whether they are likely to differ for the target population. Derive estimates using as much data as possible (e.g., when outcome data are available, use them in a principled way). The choice of method for external validity bias adjustment may be restricted by data availability (e.g., summary-level versus individual-level data) but should be driven by similar principles as those that guide the choice between outcome regressions, matching and weighting methods, and double robust approaches for confounding adjustment. Flexible nonparametric estimators that use ensemble machine learning have the potential to perform the best.

For both methods developers and applied researchers, we recommend releasing publicly available code alongside the paper and providing details for implementation. Published code facilitates replicability and accessibility of methods for future research and applied use. A substantial barrier to the adoption of new statistical methods, including advances in generalizability and transportability, is the lack of available computational tools.

While much of the causal inference literature has focused on issues of internal validity, both internal and external validity are necessary for valid inference. When treatment effect heterogeneity exists, as is often the case, study results may not hold for a target population of interest. Approaches to address internal validity biases can be borrowed to improve upon methods for addressing external validity bias. This review presents a framework for such analysis and summarizes different choices for estimators that can be used to generalize or transport results to a population different from the one under study. It brings together diverse cross-disciplinary literature to provide guidance for both applied and methods researchers. Improving the incorporation of results from observational studies can lead to better inference for decision-relevant populations with reduced bias and improved precision.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants DP2MD012722, 32LM012411, and T32ES07142. The authors thank Sebastien Haneuse, Francesca Dominici, and Laura Hatfield for feedback.

LITERATURE CITED

Ackerman B, Schmid I, Rudolph KE, Seamans MJ, Susukida R, et al. 2019. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addict. Behav.* 94:124–32

Allcott H, Mullainathan S. 2012. *External validity and partner selection bias*. NBER Work. Pap. 18373

Angrist JD, Fernández-Val I. 2013. ExtrapoLATE-ing: external validity and overidentification in the LATE framework. In *Advances in Economics and Econometrics*, Vol. 3: *Econometrics*, pp. 401–34. Cambridge, UK: Cambridge Univ. Press

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):7353–60

Attanasio O, Meghir C, Szekely M. 2003. *Using randomised experiments and structural models for 'scaling up': evidence from the PROGRESA evaluation*. IFS Work. Pap. EWP04/03, Inst. Fisc. Stud., London

Bareinboim E, Pearl J. 2014. Transportability from multiple environments with limited experiments: completeness results. *Adv. Neural Inf. Process. Syst.* 27:280–88

Bareinboim E, Pearl J. 2016. Causal inference and the data-fusion problem. *PNAS* 113(27):7345–52

Bareinboim E, Tian J. 2015. Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3475–81. Palo Alto, CA: AAAI Press

Bareinboim E, Tian J, Pearl J. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2410–16. Palo Alto, CA: AAAI Press

Begg CB. 1992. *Cross design synthesis: a new strategy for medical effectiveness research*. Rep. GAO/PEMD-92-18, US Gen. Account. Off., Washington, DC

Bell SH, Olsen RB, Orr LL, Stuart EA. 2016. Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educ. Eval. Policy Anal.* 38(2):318–35

Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, et al. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Med.* 12(10):e1001885

Bennett M, Vielma JP, Zubizarreta JR. 2020. Building representative matched samples with multi-valued treatments in large observational studies. *J. Comput. Graph. Stat.* 29(4):744–57

Buchanan AL, Hudgens MG, Cole SR, Mollan KR, Sax PE, et al. 2018. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. R. Stat. Soc. Ser. A* 181(4):1193–209

Cahan A, Cahan S, Cimino JJ. 2017. Computer-aided assessment of the generalizability of clinical trial results. *Int. J. Med. Inf.* 99:60–66

Chan W. 2017. Partially identified treatment effects for generalizability. *J. Res. Educ. Eff.* 10(3):646–69

Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. 2021. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* 4:123–44

Chen S, Tian L, Cai T, Yu M. 2017. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* 73(4):1199–209

Chipman HA, George EI, McCulloch R. 2007. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, ed. B Schölkopf, J Platt, T Hofmann, pp. 265–72. Cambridge, MA: MIT Press

Chipman HA, George EI, McCulloch RE. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4(1):266–98

Cole SR, Stuart EA. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidemiol.* 172(1):107–15

Colnet B, Mayer I, Chen G, Dieng A, Li R, et al. 2021. Causal inference methods for combining randomized trials and observational studies: a review. arXiv:2011.08047 [stat.ME]

Correa JD, Tian J, Bareinboim E. 2018. Generalized adjustment under confounding and selection biases. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 6335–42. Palo Alto, CA: AAAI Press

Cronbach LJ, Shapiro K. 1982. *Designing Evaluations of Educational and Social Programs*. Jossey-Bass Ser. Soc. Behav. Sci. High. Educ. San Francisco: Jossey-Bass. 1st ed.

Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2008. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* 90(3):389–405

Dahabreh IJ, Hernan MA, Robertson SE, Buchanan A, Steingrimsson JA. 2019a. Generalizing trial findings using nested trial designs with sub-sampling of non-randomized individuals. arXiv:1902.06080 [stat.ME]

Dahabreh IJ, Robertson SE, Petito LC, Hernán MA, Steingrimsson JA. 2022. Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *Biometrics*. In press. **https://doi.org/10.1111/biom.13716**

Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. 2020. Extending inferences from a randomized trial to a new target population. *Stat. Med.* 39(14):1999–2014

Dahabreh IJ, Robertson SE, Stuart E, Hernan M. 2017. Extending inferences from randomized participants to all eligible individuals using trials nested within cohort studies. arXiv:1709.04589 [stat.ME]

Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. 2019b. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* 75(2):685–94

Dahabreh IJ, Robins JM, Haneuse SJPA, Saeed I, Robertson SE, et al. 2019c. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. arXiv:1905.10684 [stat.ME]

Degtiar I, Layton T, Wallace J, Rose S. 2021. Conditional cross-design synthesis estimators for generalizability in Medicaid. arXiv:2109.13288 [stat.ME]

Ding P, Feller A, Miratrix L. 2016. Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B* 78(3):655–71

Dong N, Stuart EA, Lenis D, Quynh Nguyen T. 2020. Using propensity score analysis of survey data to estimate population average treatment effects: a case study comparing different methods. *Eval. Rev.* 44(1):84–108

Eddy DM. 1989. The confidence profile method: a Bayesian method for assessing health technologies. *Oper. Res.* 37(2):210–28

Fang A. 2017. 10 things to know about heterogeneous treatment effects. *EGAP, Institute of Governmental Studies*. **https://egap.org/resource/10-things-to-know-about-heterogeneous-treatment-effects/**

Flores CA, Mitnik OA. 2013. Comparing treatments across labor markets: an assessment of nonexperimental multiple-treatment strategies. *Rev. Econ. Stat.* 95(5):1691–707

Ford I, Norrie J. 2016. Pragmatic trials. *New Engl. J. Med.* 375(5):454–63

Frangakis C. 2009. The calibration of treatment effects from clinical trials to target populations. *Clin. Trials* 6(2):136–40

Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. 2009. Dealing with heterogeneity of treatment effects: Is the literature up to the challenge? *Trials* 10(1):43

Gail M, Simon R. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41(2):361–72

Gechter M. 2015. *Generalizing the results from social experiments: theory and evidence from Mexico and India*. Work. Pap., Dep. Econ., Boston Univ., Boston, MA. **https://www.bu.edu/econ/files/2015/05/Gechter_Generalizing_Social_Experiments.pdf**

Gelman A, Little TC. 1997. Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* (23):127–35

Glauner P, Migliosi A, Meira J, Valtchev P, State R, Bettinger F. 2017. Is big data sufficient for a reliable detection of non-technical losses? In *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pp. 1–6. New York: IEEE

Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511

Greenhouse JB, Kaizar EE, Anderson HD, Bridge JA, Libby AM, et al. 2017. Combining information from multiple data sources: an introduction to cross-design synthesis with a case study. In *Methods in Comparative Effectiveness Research*, ed. C Gatsonis, SC Morton, pp. 223–46. London: Chapman & Hall/CRC

Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. 2008. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Stat. Med.* 27(11):1801–13

Greenland S. 2005. Multiple-bias modelling for analysis of observational data. *J. R. Stat. Soc. Ser. A* 168:267–91

Gunter L, Zhu J, Murphy S. 2011. Variable selection for qualitative interactions. *Stat. Methodol.* 8(1):42–55

Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. 2009. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology* 32(3):229–39

Hartman E, Grieve R, Ramsahai R, Sekhon JS. 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. R. Stat. Soc. Ser. A* 178(3):757–78

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153–61

Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, et al. 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 19(6):766–79

Hill JL. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20(1):217–40

Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47(260):663–85

Hotz VJ, Imbens GW, Mortimer JH. 2005. Predicting the efficacy of future training programs using past experiences at other locations. *J. Econom.* 125(1):241–70

Imai K, King G, Stuart EA. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A* 171(2):481–502

Johansson FD, Kallus N, Shalit U, Sontag D. 2018. Learning weighted representations for generalization across designs. arXiv:1802.08598 [stat.ME]

Josey KP, Yang F, Ghosh D, Raghavan S. 2021. A calibration approach to transportability with observational data. arXiv:2008.06615 [stat.ME]

Kaizar EE. 2011. Estimating treatment effect via simple cross design synthesis. *Stat. Med.* 30(25):2986–3009

Kaizar EE. 2015. Incorporating both randomized and observational data into a single analysis. *Annu. Rev. Stat. Appl.* 2:49–72

Kallus N, Puli AM, Shalit U. 2018. Removing hidden confounding by experimental grounding. arXiv:1810.11646 [stat.ME]

Keiding N, Louis TA. 2016. Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. R. Stat. Soc. Ser. A* 179(2):319–76

Keiding N, Louis TA. 2018. Web-based enrollment and other types of self-selection in surveys and studies: consequences for generalizability. *Annu. Rev. Stat. Appl.* 5:25–47

Kern HL, Stuart EA, Hill J, Green DP. 2016. Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* 9(1):103–27

Kim JK, Park S, Chen Y, Wu C. 2018. Combining non-probability and probability survey samples through mass imputation. arXiv:1812.10694 [stat.ME]

Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. 2017. Generalizing study results: a potential outcomes perspective. *Epidemiology* 28(4):553–61

Lu Y, Scharfstein DO, Brooks MM, Quach K, Kennedy EH. 2019. Causal inference for comprehensive cohort studies. arXiv:1910.03531 [stat.ME]

Luedtke A, Carone M, van der Laan MJ. 2019. An omnibus non-parametric test of equality in distribution for unknown functions. *J. R. Stat. Soc. Ser. B* 81(1):75–99

Lunceford JK, Davidian M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 23(19):2937–60

Marcus S. 1997. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *J. Clin. Epidemiol.* 50(7):823–28

Miettinen OS. 1972. Standardization of risk ratios. *Am. J. Epidemiol.* 96(6):383–88
Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. 2017. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* 11(1):225–47
Nie L, Zhang Z, Rubin D, Chu J. 2013. Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. *Ann. Appl. Stat.* 7(3):1796–813
Olsen RB, Orr LL, Bell SH, Stuart EA. 2013. External validity in policy evaluations that choose sites purposively: external validity in policy evaluations. *J. Policy Anal. Manag.* 32(1):107–21
O'Muircheartaigh C, Hedges LV. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *J. R. Stat. Soc. Ser. C* 63(2):195–210
Pan Q, Schaubel DE. 2009. Evaluating bias correction in weighted proportional hazards regression. *Lifetime Data Anal.* 15(1):120–46
Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press
Pearl J. 2015. Generalizing experimental findings. *J. Causal Inference* 3(2):259–66
Pearl J, Bareinboim E. 2011. Transportability of causal and statistical relations: a formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, ed. M Spiliopoulou, H Wang, D Cook, J Pei, W Wang, et al., pp. 540–47. New York: IEEE
Pearl J, Bareinboim E. 2014. External validity: from do-calculus to transportability across populations. *Stat. Sci.* 29(4):579–95
Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. 2018. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med. Decis. Mak.* 38(2):200–11
Pool I, Abelson R, Popkin S. 1964. *Candidates, Issues, and Strategies: A Computer Simulation of the 1960 Presidential Election*. Cambridge, MA: MIT Press
Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, et al. 2005. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am. J. Epidemiol.* 162(5):404–14
Prevost TC, Abrams KR, Jones DR. 2000. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Stat. Med.* 19(24):3359–76
Qian M, Chakraborty B, Maiti R. 2019. A sequential significance test for treatment by covariate interactions. arXiv:1901.08738 [stat.ME]
Raudenbush SW, Schwartz D. 2020. Randomized experiments in education, with implications for multilevel causal inference. *Annu. Rev. Stat. Appl.* 7:177–208
Rothwell PM. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 365(9453):82–93
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66(5):688–701
Rudolph K, van der Laan M. 2017. Robust estimation of encouragement design intervention effects transported across sites. *J. R. Stat. Soc. Ser. B* 79(5):1509–25
Schmid I, Rudolph KE, Nguyen TQ, Hong H, Seamans MJ, et al. 2020. Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations. *Commun. Stat. Simul. Comput.* **https://doi.org/10.1080/03610918.2020.1741621**
Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c332
Schwartz D, Lellouch J. 1967. Explanatory and pragmatic attitudes in therapeutical trials. *J. Chronic Dis.* 20(8):637–48
Shadish WR, Cook TD, Campbell DT. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin
Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, et al. 2010. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics* 28(10):935–45
Simon R. 1982. Patient subsets and variation in therapeutic efficacy. *Br. J. Clin. Pharmacol.* 14(4):473–82
Stuart EA. 2010. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1):1–21
Stuart EA, Ackerman B, Westreich D. 2018. Generalizability of randomized trial results to target populations: design and analysis possibilities. *Res. Soc. Work Pract.* 28(5):532–37

Stuart EA, Bradshaw CP, Leaf PJ. 2015. Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* 16(3):475–85

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. 2011. The use of propensity scores to assess the generalizability of results from randomized trials: use of propensity scores to assess generalizability. *J. R. Stat. Soc. Ser. A* 174(2):369–86

Su X, Tsai CL, Wang H, Nickerson DM, Li B. 2009. Subgroup analysis via recursive partitioning. *SSRN Electron. J.* 10:141–58

Su X, Zhou T, Yan X, Fan J, Yang S. 2008. Interaction trees with censored survival data. *Int. J. Biostatist.* 4(1). **https://doi.org/10.2202/1557-4679.1071**

Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. 2014. A simple method for estimating interactions between a treatment and a large number of covariates. *J. Am. Stat. Assoc.* 109(508):1517–32

Tipton E. 2013a. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J. Educ. Behav. Stat.* 38(3):239–66

Tipton E. 2013b. Stratified sampling using cluster analysis: a sample selection strategy for improved generalizations from experiments. *Eval. Rev.* 37(2):109–39

Tipton E. 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* 39(6):478–501

Tipton E, Hallberg K, Hedges LV, Chan W. 2017. Implications of small samples for generalization: adjustments and rules of thumb. *Eval. Rev.* 41(5):472–505

Tipton E, Olsen RB. 2018. A review of statistical methods for generalizing from evaluations of educational interventions. *Educ. Res.* 47(8):516–24

Tipton E, Peck LR. 2017. A design-based approach to improve external validity in welfare policy evaluations. *Eval. Rev.* 41(4):326–56

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. 2009. Bias modelling in evidence synthesis. *J. R. Stat. Soc. Ser. A* 172(1):21–47

Varadhan R, Henderson NC, Weiss CO. 2016. Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogeneous: Part I. Methodology. *Commun. Stat. Case Stud. Data Anal. Appl.* 2(3–4):112–26

Verde PE, Ohmann C. 2015. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Res. Synth. Methods* 6(1):45–62

Verde PE, Ohmann C, Morbach S, Icks A. 2016. Bayesian evidence synthesis for exploring generalizability of treatment effects: a case study of combining randomized and non-randomized results in diabetes. *Stat. Med.* 35(10):1654–75

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. 2008. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* 61(4):344–49

Weisberg HI, Hayden VC, Pontes VP. 2009. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? *Clin. Trials* 6(2):109–18

Weiss CO, Segal JB, Varadhan R. 2012. Assessing the applicability of trial evidence to a target sample in the presence of heterogeneity of treatment effect. *Pharmacoepidemiol. Drug Saf.* 21:121–29

Weng C, Li Y, Ryan P, Zhang Y, Liu F, et al. 2014. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl. Clin. Inform.* 5(2):463–79

Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. 2017. Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* 186(8):1010–14