Current Epidemiology Reports (2020) 7:117–124
https://doi.org/10.1007/s40471-020-00239-0

EPIDEMIOLOGIC METHODS (P HOWARDS, SECTION EDITOR)

Target Validity: Bringing Treatment of External Validity in Line with Internal Validity

Catherine R. Lesko[1] · Benjamin Ackerman[2] · Michael Webster-Clark[3] · Jessie K. Edwards[3]

Published online: 30 June 2020

Abstract

Purpose of Review "Target bias" is the difference between an estimate of association from a study sample and the causal effect in the target population of interest. It is the sum of internal and external bias. Given the extensive literature on internal validity, here, we review threats and methods to improve external validity.

Recent Findings External bias may arise when the distribution of modifiers of the effect of treatment differs between the study sample and the target population. Methods including those based on modeling the outcome, modeling sample membership, and doubly robust methods are available, assuming data on the target population is available.

Summary The relevance of information for making policy decisions is dependent on both the actions that were studied and the sample in which they were evaluated. Combining methods for addressing internal and external validity can improve the policy relevance of study results.

Keywords External validity · Generalizability · Internal validity · Randomized trials · Target population · Transportability

Introduction

When identifying the most relevant information for policy makers or clinicians looking to make a decision about how to act in a particular population or for a particular patient, both the actions being considered and the context to which they will be applied matter [1, 2]. One hierarchy of study designs places results from randomized controlled trials (RCTs) at the pinnacle of the pyramid of evidence because RCTs minimize *internal bias* due to confounding by design through randomization [3]. Setting aside the fact that RCTs may still suffer from internal biases other than confounding bias, RCTs often are conducted in highly selected study samples that may yield a very different context than the target population in which the decision is being made. This mismatch of context in and composition of the trial sample and the target population is a key component of *external bias*, which is undervalued in this evidence hierarchy.

Lest we forget how much the target population matters, we present two examples: (1) estimates of the effect of medication assisted therapy (buprenorphine/naloxone), motivational interviewing, and motivational incentives on substance use would have been very different—typically less effective and no longer statistically significant—had trials testing these interventions been conducted in samples that were more representative of all treatment-eligible persons in the US [4]. (2) Estimates of the (adverse) effect of antidepressants on suicidal ideation or behaviors in depressed youth may have been overstated in trials that under-enrolled or explicitly excluded youth at the highest risk for these outcomes [5, 6].

The term *target validity* has been proposed to describe the total difference between the estimate of association obtained in a particular study sample and the true effect in the target population of interest [7••]. Target bias is the sum of internal bias and external bias. We loosely define internal bias as the difference between the estimate of association in the study sample and the true effect in the study sample and external bias as the difference between the true effect in the study

This article is part of the Topical Collection on *Epidemiologic Methods*

✉ Catherine R. Lesko
clesko2@jhu.edu

[1] Department of Epidemiology, Johns Hopkins School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, USA

[2] Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, USA

[3] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA

sample and the true effect in the target population. Although the moniker "target validity" is new, the concept has been previously described in the education, social sciences, and policy literature [1, 6, 8, 9]. The concept of target validity encourages epidemiologists to take threats to external validity as seriously as they have traditionally taken threats to internal validity, to integrate consideration of both internal and external validity when evaluating the strength of available evidence for informing a particular decision in a particular population, and to better evaluate the tradeoffs between experimental and observational studies [9]. This is in contrast to the common view that external validity is secondary to, or contingent on, internal validity [10].

Although threats to internal validity are well-known to epidemiologists (e.g., confounding bias, information bias), threats to external validity are less well-understood. Focusing on the external validity of RCTs is a beneficial heuristic in that (1) the sampling mechanism into a trial, including inclusion and exclusion criteria, is often explicit and thus so are differences between the trial sample and the target population; (2) we know quite a bit about the implicit sampling mechanisms into trials (e.g., under-sampling of older people and people of minority race/ethnicity) thanks to previous research [11–14]; and (3) we can pretend that bias due to confounding negligible and can thus assume that the majority of the differences (bias) between the estimate of association in the sample and the target (population) average treatment effect (TATE) of interest is due to lack of external validity. That is, we assume the estimate of association in the sample is a good estimate of the sample average treatment effect (SATE), but the SATE is a poor approximation of the TATE. Despite the relative inattention paid to external validity, the assumptions required for external validity may be quite strong [15•].

Despite the specter of unmeasured confounding in observational studies, they have generally (despite some notable exceptions) returned similar results as subsequent randomized trials investigating the same exposures [16]. That is, for many exposures, particularly those that *could* be randomized, internal validity of observational studies may be better than is often assumed. Additionally, observational studies do not tend to have as strict of inclusion and exclusion criteria as trials, making them potentially more similar to the target populations we might be interested in. However, the external validity of observational studies is still potentially of concern, given the increasing use of "big data," administrative databases, and pooled or collaborative cohort studies, which rely on samples that arise from sampling mechanisms that are myriad and unclear [17].

Formal Frameworks for External Validity

Threats to target validity associated with internal validity (e.g., confounding bias) have been extensively described. We define internal bias as the difference between the association measured in the sample, $E[Y|A=1, S=1]-E[Y|A=0, S=1]$, and the SATE, $E(Y^{a=1}-Y^{a=0}|S=1)$. Here, we use Y to denote the outcome, A to denote treatment received, $S=1$ to denote membership in the study sample, and Y^a to denote the outcome Y that would occur if treatment a were assigned (the potential outcome).

Next, we informally define and describe key threats to external validity, since these threats have been less frequently explored in the literature. There are at least three reasons that the SATE may not equal the TATE (we will wait to formally define this latter quantity until later in this section, for reasons that will become clear), including:

- There are modifiers, Z, of the effect of treatment, and the distribution of those modifiers is different in the study sample and the target population [1, 8, 18–20, 21•, 22–24];
- The version of treatment (including details of how the treatment is delivered) impacts the effect of treatment, and the distribution of the versions of treatment is different in the study sample and the target population [25]; and
- There is interference (one persons' exposure impacts another persons' outcome), and the patterns of interference differ between the two populations.

It is (typically implicitly) assumed that sample membership or trial participation itself, $S=1$, does not have a direct effect on the outcome [21•]; that is, if sample membership was itself an intervention (e.g., if the act of being observed as part of being in the study changes participants behavior in a way that changes the outcome not directly through receipt of the intervention), the "versions" of treatment in the study sample and the target population would differ, and reason 2 above would lead to differences between the SATE and the TATE [25].

The majority of work done on external validity of study results has focused on differences in the distribution of effect modifiers—that is, external biases related to sample composition. The magnitude of the external biases related to sample composition is a function of the probability of selection into the sample, the heterogeneity of treatment effects, and the association between sample membership and effect modifiers [1, 18, 24]. Existing frameworks for describing this problem, determining identifiability of the TATE, and defining estimators of the TATE are more similar to one another than they are different. A key feature of these all of these frameworks for defining external validity, however, is that the target population needs to be well-characterized (theoretically enumerable). Moving forward, for mainly logistic but sometimes theoretical reasons, we split out external validity into "generalizability" and "transportability."

"Generalizability" refers to the situation in which our study sample is a proper subset of the target population, but the study sample may or may not be a simple random sample from the target population. That is, the TATE of interest is $E(Y^{a=1}$

$-Y^{a=0}$), or the effects in a target population of which the study sample are members. If the study sample is a simple random sample of the target population, results from that study are generalizable to the target population in expectation. Multiple methods are available to adjust for the situation in which the study sample is not a simple random sample of the target population under two key assumptions: mean exchangeability between the sample and the target population, $E[Y^a|Z, S=1] = E[Y^a|Z]$ for every $a \in \mathcal{A}$; and positivity of trial participation, $P(S=1|Z=z) > 0$ for all z such that $P(Z=z) > 0$ [21•, 22]. Again, Z is the set of modifiers that are associated with sample membership; however, it is worth noting that all covariates that are associated with the outcome will be effect measure modifiers on at least one scale. The positivity assumption implies that *tractable* generalizability problems (i.e., situations in which estimation of a TATE is possible from a study sample because identifiability can conceivably be met) are those that do not require extrapolation beyond the characteristics of the persons in the study sample [21•]. The practical implication of the positivity assumption is that in order for trials to hope to provide good information about the expected TATE, they must enroll a full spectrum of patients; for example, trials conducted only in adults < 50 years old cannot provide information about the effect of interest in a target population that includes all adults without making the (strong) assumption that age does not modify that effect.

In contrast, "transportability" has been used to refer to the situation in which our study sample and target population are not overlapping [25–27]. That is, the TATE of interest is $E(Y^{a=1} - Y^{a=0}|S=0)$; we are interested in the effect in a set of persons who were NOT members of the study sample or the complement to the study sample in the set of individuals created by combining the study sample and the target population. Transportability, then, *may* involve extrapolation to a different context [26]. Another definition of transportability that has been put forward is the "extension of inferences from [a] trial to a target population that includes participants who are not part of the trial-eligible population" [28]. Here, the implication is that the TATE of interest is $E(Y^{a=1} - Y^{a=0})$, but the positivity assumption is not met. This definition of transportability explicitly involves extrapolation beyond the characteristics of the persons in the study. Transportability, then, appears to require stronger assumptions than generalizability from a theoretical sense.

Confusingly, the term "transportability" (or "transportability weights") has also been used occasionally to describe *methods* for extending inference from a study sample to a target population to which the study sample belongs (a "generalizability" problem, as defined above), when data on the entire target population is unavailable or the particular subset of the target that participated in the study is not enumerable [27]. For example, we may have a trial of antihypertensive treatment conducted in a sample of adults with hypertension living in the United States (US) but no data on the full target population (all adults with hypertension living in the US). Instead, we may have data on a random or representative sample of the target population. Large governmental surveys often serve this purpose (e.g., NHANES). In this situation, we do not want to "generalize" to the population represented by the union of the data from the study sample, denoted $S^* = 1$, and the target population (or a sample of the target population), denoted $S^* = 0$. The TATE of interest is no longer $E[Y^{a=1} - Y^{a=0}]$, but rather it is $E[Y^{a=1} - Y^{a=0}|S^* = 0]$. Thus, sometimes even when we are theoretically generalizing results, we might use methods that were designed for "transportability" [19].

Assumptions and Identifiability

A set of assumptions sufficient to identify the TATE parallel a sufficient set of assumptions for identification of the SATE. In addition to mean exchangeability between the sampled and unsampled members of the target population perhaps conditional on covariates and positivity, we assume treatment version irrelevance (also called causal consistency) or the same distribution of versions of treatment in the study sample and the target population; no interference or the same patterns of interference in the study sample and the target population; no measurement error including of all Z variables; and correct causal model(s) specification [22].

Barenboim and Pearl proposed the use of "selection diagrams," an extension of directed acyclic graphs (DAGs), for encoding assumptions about causal relationships in the sample and in a distinct target populations and then determining whether a TATE is identifiable from the available data [26]. As long as the characteristics that differ between the two populations are all pre-treatment covariates, the assumptions sufficient for generalizability and transportability, and sufficient sets of covariates for mean exchangeability between the sample and the target population, coincide [29].

Assessing the Generalizability or Transportability of Effects

Important limitations of existing study results for guiding policy or treatment decisions as related to inclusion of key populations in public health and medical research have been qualitatively recognized for some time [14, 30–38]. Quantitative assessments of the differences between a study sample and target population improve the rigor of such exercises and include, for example, reweighting the sample by the inverse probability of membership in the sample and then comparing the differences in characteristics of the weighted sample and the target population using standardized mean differences [39, 40]. However, for any (qualitatively or quantitatively observed) differences in sample composition to result in external bias, those characteristics that differ between the study sample and target population(s) of interest must also modify the treatment effect [5].

Any predictors of the outcome are likely to modify the treatment effect on at least one scale. This implies that assessing the generalizability or transportability of effects not only requires specifying the target population to whom one would like to make inference but also the scale on which one would like to report results. There are mathematical arguments supporting the idea that odds ratios are the least heterogeneous measure of association, while risk differences are most heterogeneous; however, absolute measures of effect are arguably most meaningful from both a public health and etiologic perspective.

Methods to Account for External Bias

In Design

The best way to ensure target validity in expectation would be to randomly sample the study sample from the target population (ensuring external validity in expectation), then randomly assign treatment to members of the study sample. Random sampling of the target population would ensure the sample is representative of the target population on both measured and unmeasured covariates, in expectation. However, random sampling of the target population is often not possible for logistical, ethical, or practical reasons [41–44]. One design option to improve the generalizability of trial results is purposive stratified sampling [45, 46] or pragmatic or practical clinical trials that tend to have less restrictive inclusion and exclusion criteria [47, 48]. However, while pragmatic trials these are more likely to be generalizable than traditional efficacy trials, they are still not expected to yield study samples that perfectly reflect the target population. Furthermore, many questions about generalizability and transportability of study arise after the research has been conducted with reference to a new target population. It is far more efficient to use existing study results to estimate or approximate the TATE in each of these new target populations than it would be to conduct new, separate trials in all possible target populations of interest.

In Analysis

Just as methods exist to account for non-random treatment assignment (e.g., regression adjustment, propensity score methods including weighting, g-computation or standardization, and doubly robust methods), methods exist to account for non-random sampling into the study sample. Most of these methods are analogous to those used to account for confounding and selection bias. Broadly, these methods can be grouped into methods based on modeling the probability of the outcome, methods based on modeling the probability of sample membership, and doubly robust methods that combine the two approaches [21•, 24, 49, 50].

Outcome model-based approaches to account for sample composition of a trial typically involve estimating subgroup-specific treatment effects from the sample and the averaging them by the proportion of the target population in each of those subgroups. Pearl termed this the "post-stratification formula." Fundamentally, the formula looks just like Robins's g-formula [51], where rather than estimating the average treatment effect by averaging over the distribution of covariates in the sample, we average over the distribution of covariates in the target population. One outcome model-based approach to generalizing study results is to model the outcome as a flexible function of observed covariates using data from the study sample and then predict outcomes for all members of the target population under each treatment of interest. This has been done using parametric regression models [52] and machine learning methods [53, 54].

Alternatively, the treatment effect can be estimated in the study sample that has been weighted to look like the target population. Specific details regarding the construction of these sample membership weights depends on whether we envision the problem as one of transportability or generalizability, both theoretically, but also practically. If we have a dataset enumerating the target population and also the specific members of the target population who were selected into the study sample (a generalizability problem both theoretically and practically), the weights are simply the inverse of the probability of sample membership for everyone in the sample and zero for everyone else. If, on the other hand, the study sample is not a subset of the target population (a transportability problem theoretically) or data on the target population and the study sample are not linkable (a transportability problem practically), we turn to a different set of weights. Data on the study sample and the target population may not be linkable if we only have data on a sample of the target population (e.g., from a sample survey), or we do not have data on which members of the target population were included in the sample (e.g., from administrative records) [55•]. "Transportability" weights are the inverse of the odds of sample membership for everyone in the sample and zero for everyone else [19, 27].

In both generalizability and transportability problems, weights are typically estimated using predicted probabilities from a sample membership model, with generalizability weights akin to propensity score weighting (inverse probability of exposure weighting) and transportability weights resembling those for estimating the average treatment among the treated (ATT) [56]. Applications of sample membership weighting methods have been most prevalent in the literature (relative to outcome model-based methods) [18, 24, 57].

In order to implement the methods described above, one must find an appropriate secondary data source on the target population of interest, which can be quite challenging to do in practice [58]. First, the data must include comparable measures on a sufficient set of covariates, such that the assumption

of mean exchangeability between the sample and the target is met. Sensitivity analyses are possible if data on effect modifiers are missing in the target population, or in both data sets, for example by specifying plausible distributions of those effect modifiers and the strength of the effect modification [59, 60]. Second, one must assume that the population data are either a random sample of the true target population of interest or a complete census. However, many promising sources of publicly available data come from complex surveys, like the National Survey on Drug Use and Health (NSDUH), where it is known that study participants were not in fact randomly sampled [61]. Simply transporting RCT results to survey samples like NSDUH without properly accounting for the sampling methodology will result in biased estimates of the TATE. In other words, generalizations may be accurate for the NSDUH study sample but not for the population that the NSDUH sample represents. Recent methodological work has addressed this by determining how to incorporate survey weights from population data into existing generalizability methods [62].

Novel Study Designs to Account for External Bias

Two particularly novel study designs have been demonstrated in real-world data to account for external bias and explicitly assess the plausibility of some of the key assumptions (detailed above).

Nesting Trials Within Clinical Cohorts

If trials obtain permission from participants to link to their medical record data or if trials are nested within medical systems such that trial participants are identifiable within the population of patients who would have been eligible to participate in the trial, there are unique study design possibilities. Most basic is the potential for the methods described above to be used to generalize trial results to the broader target population [19, 22, 28, 49]. Alternatively, if the treatment under study is available outside of the trial, trial results that have been generalized to the cohort could be compared with the estimated effect of treatment in the cohort based on observational data [63]. Generalized trial results would be expected to differ from the truth if the adjustment set did not include a sufficient set of modifiers, while the association between exposure and outcome estimated in the target population directly using non-randomized treatment would be expected to differ from the truth if the adjustment set did not include a sufficient set of confounders [63–67]. If results from the two approaches are similar, we can have more confidence in the estimate of the TATE [21•, 63]. This is an example of triangulation of study results [68].

Leveraging Lack of Treatment Availability Outside a Trial

If at least one arm of the trial (treatment $A = a$) is currently available in the target population, the assumption of "mean generalizability" or "mean exchangeability over S" [21•] can be partially evaluated by comparing generalizing the outcomes from the trial under treatment to observed outcomes under treatment a in the target population [69]. A major difference between the generalized and observed outcomes implies that the generalized treatment effect for treatment a versus a' is likely to differ from the true TATE. This type of analysis is particularly useful when studying a novel treatment where the placebo or standard of care arm of the trial is the only (currently) available treatment in the target population. In such cases, comparing the generalized outcomes in the placebo arm of the trial to the observed outcomes in the target population gives a sense of whether the (partial) conditional exchangeability assumption over S is likely to have been met. This approach has been demonstrated in education [24] and health using contemporaneous controls in the general population before broad availability of the treatment under study [70]. It has also been demonstrated using historical controls when estimating the effects of experimental medical treatments for terminally ill patients available under "right to try" laws [71]. Critically, this method only tests for a *failure* to transport or generalize. Even if the generalized outcomes and observed outcomes in the target population are identical under a, there is no guarantee that the generalized outcomes under treatment a' will equal the unobserved outcomes in the target population under a'.

Conclusions

The utility of an estimate for informing a public health decision is a function of how accurately it maps to the causal effect of interest in the relevant target population [2]. Recent work on target validity has focused on increasing awareness of impact that external bias has on overall bias. Valid estimates of effect for relevant target populations are attainable given rich descriptive data on target populations, and new methods for extending results from one study sample to another population, under the set of assumptions described above. The strength of assumptions under which such an extension is possible is a function of how different the study sample and target population are from one another with respect to covariates that modify the effect of treatment. Although representativeness (of a study sample with respect to the target population) may not be necessary for all studies [41–44], the distribution of covariates in the sample and the target population should, at a very

minimum, be considered when answering policy-relevant questions. Methods are available to account for differences in measured covariates between a study sample and target population and should be carefully implemented when drawing population inferences from non-representative samples.

Funding Information This work was supported by National Institutes of Health grants K01 AA028193 and K01 AI125087 and US Department of Education Institution of Education Sciences grant R305D150003.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent All reported studies/experiments with human or animal subjects performed by the authors have been previously published and complied with all applicable ethical standards (including the Helsinki declaration and its amendments, institutional/national research committee standards, and international/national/institutional guidelines).

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Olsen RB, Orr LL, Bell SH, Stuart EA. External validity in policy evaluations that choose sites purposively. J Policy Anal Manag. 2013;32(1):107–21.
2. Edwards JK, Lesko CR, Keil AP. Invited commentary: causal inference across space and time-quixotic quest, worthy goal, or both? Am J Epidemiol. 2017;186(2):143–5.
3. US Preventive Services Task Force United States. Office of Disease Prevention Health Promotion. Guide to clinical preventive services: report of the US Preventive Services Task Force: US Department of Health and Human Services, Office of Public Health and …; 1996.
4. Susukida R, Crum RM, Ebnesajjad C, Stuart EA, Mojtabai R. Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. Addiction. 2017;112(7):1210–9.
5. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. Stat Med. 2008;27(11):1801–13.
6. Weisberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? Clinical Trials. 2009;6(2):109–18.
7.•• Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target validity and the hierarchy of study designs. Am J Epidemiol. 2019;188(2):438–43 **Introduces the term target validity and defines target validity both formally and conceptually.**
8. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2001. xxi, 623 p. p.
9. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc a Stat. 2008;171:481–502.
10. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. x, 758 p. p.
11. Curno MJ, Rossi S, Hodges-Mameletzis I, Johnston R, Price MA, Heidari S. A systematic review of the inclusion (or exclusion) of women in HIV research: from clinical studies of antiretrovirals and vaccines to cure strategies. J Acquir Immune Defic Syndr. 2016;71(2):181–8.
12. Geller SE, Koch AR, Roesch P, Filut A, Hallgren E, Carnes M. The more things change, the more they stay the same: a study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. Acad Med. 2018;93(4):630–5.
13. Green BL, Maisiak R, Wang MQ, Britt MF, Ebeling N. Participation in health education, health promotion, and health research by African Americans: effects of the Tuskegee Syphilis Experiment. J Health Educ. 1997;28(4):196–201.
14. Susukida R, Crum RM, Stuart EA, Ebnesajjad C, Mojtabai R. Assessing sample representativeness in randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. Addiction. 2016;111(7):1226–34.
15.• Breskin A, Westreich D, Cole SR, Edwards JK. Using bounds to compare the strength of exchangeability assumptions for internal and external validity. Am J Epidemiol. 2019;188(7):1355–60 **Derives nonparametric bounds to compare strengths of assumptions for internal and external validity under which a TATE (causal risk difference) can be estimated from an associational risk difference.**
16. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92.
17. Lesko CR, Jacobson LP, Althoff KN, Abraham AG, Gange SJ, Moore RD, et al. Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. Int J Epidemiol. 2018;47(2):654–68.
18. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. Am J Epidemiol. 2010;172(1):107–15.
19. Ackerman B, Schmid I, Rudolph KE, Seamans MJ, Susukida R, Mojtabai R, et al. Implementing statistical methods for generalizing randomized trial findings to a target population. Addict Behav. 2019;94:124–32.
20. Buchanan AL, Hudgens MG, Cole SR, Mollan K, Sax PE, Daar ES, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. Technical Report Series 2015.
21.• Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernan MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics. 2019;75(2):685–94 **Summary of methods and review of conditions for estimating a TATE from RCT data**.
22. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology. 2017;28(4):553–61.
23. Stuart EA, Ackerman B, Westreich D. Generalizability of randomized trial results to target populations: design and analysis possibilities. Res Soc Work Pract. 2018;28(5):532–7.
24. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. Prev Sci. 2015;16(3):475–85.
25. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011;22(3):368–77.

26. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. J Causal Inference. 2013;1(1):107–34.
27. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2017;186(8):1010–4.
28. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. Eur J Epidemiol. 2019;34(8):719–22.
29. Pearl J. Generalizing experimental findings. J Causal Inference. 2015;3(2):259–66.
30. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82–93.
31. Rothwell PM. Commentary: external validity of results of randomized trials: disentangling a complex concept. Int J Epidemiol. 2010;39(1):94–6.
32. Dekkers OM, von Elm E, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol. 2010;39(1):89–94.
33. Braslow JT, Duan N, Starks SL, Polo A, Bromley E, Wells KB. Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. Psychiatric services (Washington, DC). 2005;56(10):1261–8.
34. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA - J Am Med Assoc. 2007;297(11):1233–40.
35. Sokka T, Pincus T. Eligibility of patients in routine care for major clinical trials of anti–tumor necrosis factor α agents in rheumatoid arthritis. Arthritis Rheum. 2003;48(2):313–8.
36. Hoertel N, Le Strat Y, Blanco C, Lavaud P, Dubertret C. Generalizability of clinical trial results for generalized anxiety disorder to community samples. Depress Anxiety. 2012;29(7):614–20.
37. Blanco C, Olfson M, Okuda M, Nunes EV, Liu S-M, Hasin DS. Generalizability of clinical trials for alcohol dependence to community samples. Drug Alcohol Depend. 2008;98(1–2):123–8.
38. Steg PG, López-Sendón J, de Sa EL, Goodman SG, Gore JM, Anderson FA, et al. External validity of clinical trials in acute myocardial infarction. Arch Intern Med. 2007;167(1):68–73.
39. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A Stat Soc. 2011;174(2):369–86.
40. Tipton E. How generalizable is your experiment? An index for comparing experimental samples and populations. J Educ Behav Stat. 2014;39(6):478–501.
41. Ebrahim S, Davey SG. Commentary: should we always deliberately be non-representative? Int J Epidemiol. 2013;42(4):1022–6.
42. Richiardi L, Pizzi C, Pearce N. Commentary: representativeness is usually not necessary and often should be avoided. Int J Epidemiol. 2013;42(4):1018–22.
43. Rothman K, Hatch E, Gallacher J. Representativeness is not helpful in studying heterogeneity of effects across subgroups. Int J Epidemiol. 2014;43(2):633–4.
44. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42(4):1012–4.
45. Tipton E. Stratified sampling using cluster analysis: a sample selection strategy for improved generalizations from experiments. Eval Rev. 2013;37(2):109–39.
46. Tipton E, Hedges L, Vaden-Kiernan M, Borman G, Sullivan K, Caverly S. Sample selection in randomized experiments: a new method using propensity score stratified sampling. J Res Educ Eff. 2014;7(1):114–35.
47. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA - J Am Med Assoc. 2003;290(12):1624–32.
48. Insel TR. Beyond efficacy: the STAR* D trial. Am J Psychiatr. 2006;163(1):5–7.
49. Dahabreh IJ, Robertson SE, Hernan MA. On the relation between G-formula and inverse probability weighting estimators for generalizing trial results. Epidemiology. 2019.
50. Dahabreh IJ, Hernan MA, Robertson SE, Buchanan A, Steingrimsson JA. Generalizing trial findings in nested trial designs with sub-sampling of non-randomized individuals. arXiv preprint arXiv:190206080. 2019.
51. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. Mathe Model. 1986;7:1393–512.
52. Wang C, Mollan KR, Hudgens MG, Tucker JD, Zheng H, Tang W, et al. Generalisability of an online randomised controlled trial: an empirical analysis. J Epidemiol Community Health. 2018;72(2):173–8.
53. Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. J Res Educ Eff. 2016;9(1):103–27.
54. Rudolph KE, Schmidt NM, Glymour MM, Crowder R, Galin J, Ahern J, et al. Composition or context: using transportability to understand drivers of site differences in a large-scale housing experiment. Epidemiology (Cambridge, Mass). 2018;29(2):199.
55.• Bonander C, Nilsson A, Bergström GM, Björk J, Strömberg U. Correcting for selective participation in cohort studies using auxiliary register data without identification of non-participants. Scand J Public Health. 2019:1403494819890784. **Summary of practical issues when estimating sampling weights to generalize trial results to a target population, based on what data are available in the target population.**
56. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003;14(6):680–6.
57. Bonander C, Nilsson A, Björk J, Bergström GM, Strömberg U. Participation weighting based on sociodemographic register data improved external validity in a population-based cohort study. J Clin Epidemiol. 2019;108:54–63.
58. Stuart EA, Rhodes A. Generalizing treatment effect estimates from sample to population: a case study in the difficulties of finding sufficient data. Eval Rev. 2017;41(4):357–88.
59. Nguyen TQ, Ackerman B, Schmid I, Cole SR, Stuart EA. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: assumptions, models, effect scales, data scenarios, and implementation details. PLoS One. 2018;13(12):e0208795.
60. Hong JL, Jonsson Funk M, LoCasale R, Dempster SE, Cole SR, Webster-Clark M, et al. Generalizing randomized clinical trial results: implementation and challenges related to missing data in the target population. Am J Epidemiol. 2018;187(4):817–27.
61. Batts K, Pemberton M, Bose J, Weimer B, Henderson L, Penne M, et al. Comparing and evaluating substance use treatment utilization estimates from the National Survey on Drug Use and Health and other data sources. CBHSQ Data Review Rockville, MD: Substance Abuse and Mental Health Services Administration (US). 2014.
62. Ackerman B, Lesko CR, Siddique J, Susukida R, Stuart EA. Generalizing randomized trial findings to a target population using complex survey population data. arXiv preprint arXiv:200307500. 2020.
63. Marcus SM. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. J Clin Epidemiol. 1997;50(7):823–8.
64. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.
65. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–9.

66. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. Epidemiologic perspectives & innovations : EP+I. 2009;6:4.
67. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Stat Med. 2013;32(20):3552–68.
68. Lawlor DA, Tilling K, Davey SG. Triangulation in aetiological epidemiology. Int J Epidemiol. 2016;45(6):1866–86.
69. Hartman E, Grieve R, Ramsahai R, Sekhon JS. 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 Series A (Statistics in Society). 2015;178(3):757–78.
70. Webster-Clark MA, Sanoff HK, Stürmer T, Lund SPHJL. Diagnostic assessment of assumptions for external validity: an example using data in metastatic colorectal cancer. Epidemiology (Cambridge, Mass). 2019;30(1):103.
71. Hazlett C. Estimating causal effects of new treatments despite self-selection: the case of experimental medical treatments. Journal of Causal Inference. 2019;7(1).