Generalizing Study Results

A Potential Outcomes Perspective

Catherine R. Lesko,[a,b] *Ashley L. Buchanan,*[c,d] *Daniel Westreich,*[a] *Jessie K. Edwards,*[a] *Michael G. Hudgens,*[c] *and Stephen R. Cole*[a]

Abstract: Great care is taken in epidemiologic studies to ensure the internal validity of causal effect estimates; however, external validity has received considerably less attention. When the study sample is not a random sample of the target population, the sample average treatment effect, even if internally valid, cannot usually be expected to equal the average treatment effect in the target population. The utility of an effect estimate for planning purposes and decision making will depend on the degree of departure from the true causal effect in the target population due to problems with both internal and external validity. Herein, we review concepts from recent literature on generalizability, one facet of external validity, using the potential outcomes framework. Identification conditions sufficient for external validity closely parallel identification conditions for internal validity, namely conditional exchangeability; positivity; the same distributions of the versions of treatment; no interference; and no measurement error. We also require correct model specification. Under these conditions, we discuss how a version of direct standardization (the g-formula, adjustment formula, or transport formula) or inverse probability weighting can be used to generalize a causal effect from a study sample to a well-defined target population, and demonstrate their application in an illustrative example.

(*Epidemiology* 2017;28: 553–561)

Epidemiology as a discipline seeks to identify causes of disease for the purpose of intervening to improve public health. Great care is generally taken in epidemiologic studies to ensure the internal validity of causal effect estimates.[1] However, the external validity of effect estimates has received considerably less attention. Although there have been recent advances in methods for drawing externally valid inferences, particularly in statistics and computer science,[2–6] those concepts have not yet been widely accepted in epidemiology[7,8] as is evidenced by ongoing debates as to the importance of representativeness in study samples.[9–11] The purpose of this article is to review recent developments in generalizability, one facet of external validity, using the potential outcomes framework.

For the purposes of this article, external validity refers to the extent to which an internally valid effect measured in a study sample is an (asymptotically) unbiased estimator of the treatment effect in the population of interest (henceforth, the target population).[7] The effect in the study sample is sometimes called the sample average treatment effect, whereas the effect of interest is sometimes called the (target) population average treatment effect. External validity can be divided into two problems: generalizability and transportability. Generalizability is concerned with making inference from a possibly biased sample of the target population back to the full target population (including the study sample), whereas transportability concerns making inference for a target population when the study sample and the target population are partially or completely nonoverlapping. Here we will mainly discuss generalizability, as in references 5, 7, 12. A set of identification assumptions sufficient for generalizability are outlined, and parallels are noted with identification assumptions sufficient for internal validity. We describe two estimators that have been proposed to generalize results from a study sample to a target population when the study sample was not randomly sampled from the target population: a generalization of the g-formula[13] (adjustment formula)[14,15] and an inverse probability of sampling weighted estimator. We demonstrate their use with an illustrative example. We briefly outline distinctions between generalizability and transportability. Finally, we discuss practical considerations for addressing generalizability in epidemiologic study design.

Editor's Note: A commentary on this article appears on p. 562.
Submitted 9 February 2016; accepted 23 March 2017.

From the [a]Department of Epidemiology, University of North Carolina, Chapel Hill, NC; [b]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; [c]Department of Biostatistics, University of North Carolina, Chapel Hill, NC; and [d]Department of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA.

Funding for this study was provided in part by the National Institutes of Health (DP2 HD084070, R01 AI100654, P30 AI50410, U01 HL121812, U01 AI103390-02, R24 AI067039). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Code for the example presented in this article is available on request.
The authors report no conflicts of interest.

Correspondence: Catherine R. Lesko, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205. E-mail: clesko2@jhu.edu.

DEFINITIONS AND CAUSAL FRAMEWORK

Specifying a well-defined causal question starts by defining: the outcome(s) of interest Y; the treatments of interest A; and the target population.[15] The effect of interest (population average treatment effect) is a contrast of the average potential outcome in the target population under two different interventions, treatments, or policies of interest, for example:

$$E[Y(a)]-E[Y(a')] \tag{1}$$

where $Y(x)$ denotes the outcome that a participant would have if he or she received treatment $x=a,a'$. Unless otherwise noted, E and P denote expectations and probabilities in the target population. Equation (1) can be expanded, by the law of total probability, to incorporate pretreatment covariates W:

$$E\{E[Y(a)\mid W]-E[Y(a')\mid W]\} \;= E\{E[Y(a)-Y(a')\mid W]\} \tag{2}$$

When the effect of interest is heterogeneous over strata defined by W, Equation (2) emphasizes that the population average treatment effect is a weighted average of the stratum-specific effects, with weights defined by the distribution of W in the target population $P(W=w)$. (Here, we assume variables in W are discrete; however, all concepts are easily extended to incorporate continuous W.) Answering the causal question requires data, for example, a study sample. Although it is often assumed for statistical inference that the study sample is a random sample of the target population, such random sampling rarely occurs in practice. If the stratum-specific effects differ and the distribution of W in the study sample differs from the distribution of W in the target population, the sample average treatment (or causal) effect

$$SATE=\frac{1}{n}\sum_{i=1}^{n}[Y_i(a)-Y_i(a')]$$

(where i indexes the n individuals in the study sample, $i = 1, ..., n$) may not equal the population average treatment effect. It is helpful to distinguish threats to validity that arise after enumeration of the study sample, which we define as threats to internal validity, from threats to validity due to eli-gibility and enrollment of study subjects, which we define as threats to external validity. We define an estimator as inter-nally valid when the estimator of association in the study sample is an unbiased estimator of the sample average treat-ment effect or the average treatment effect in the population of which the study sample constitutes a simple random sam-ple (although this latter population is usually hypothetical as random sampling to generate a study sample is rare in public health research). We define a causal estimator to be externally valid when it is an unbiased estimator of the average treatment effect in the target population.

DEFINING THE TARGET POPULATION

Ideally, the overall study goal would drive the choice of the target population, the study sample would be randomly sampled from that target population, and exposure would be randomly assigned within the study sample such that the sample average treatment effect would equal the population average treatment effect in expectation. However, typically, an investigator has a causal question and a study sample within which to conduct analysis. After the research has been conducted, the investigator would like to know whether their results are "generalizable" to the population from which the study sample was drawn or "transportable" to an external target population.

Generalizability is a characteristic of the relationship between results from a specific study sample and a specific target population, not a characteristic of a study alone. Therefore, to make meaningful inference about the generalizability of study results, the target population of interest must be well defined.[9,16–19] Study results may be generalizable to one specific target population but not another. Comparisons between the target population and study sample should consider differences in patient characteristics (exchangeability), details of the intervention (treatment versions), and patterns of interference.

There are several types of target populations that may be of interest,[15] which can be classified according to their relationship with the sample. First, when the target population is not explicitly described, the implicit assumption is that the target population of interest is the study sample itself, or the hypothetical population from which the study sample was randomly sampled; this population is described by a (typical) paper's so-called Table 1. This is a commonly assumed target population in statistical and causal inference. However, it is almost never the case that the study sample is a simple random sample or census of any target population of substantive interest. Often research is conducted to inform decisions about a population at least somewhat different than that under study, or at the very least to inform decisions about the same population in the future. Second, we may be interested in a target population from which the study sample was sampled, but where the sampling was not at random. In this case, the sample average treatment effect will typically differ from population average treatment effect. In statistics and economics, this difference has been called "sampling selection bias."[3,4,20,21] Finally, we may be interested in a target population that is distinct from the study sample. If the study sample is neither a census nor a (possibly biased) sample from the target population, we face a problem of transportability rather than generalizability,[2] and additional assumptions or information are needed to estimate the effect of interest in the target population.

ASSUMPTIONS

Determining a set of assumptions sufficient to identify a causal parameter applicable to a particular set of individuals is a fundamental step in the process of causal inference. Identification involves writing a well-defined function of the distribution of potential outcomes in terms of a well-defined function of the distribution of the observed data. Within a study sample, the fundamental problem of causal inference can be framed as a missing data problem: we never observe all potential outcomes for subjects in our study sample and thus assumptions are required for parameter identification.[22,23] Sufficient sets of assumptions are well described in the literature for identification of a sample average treatment effect (i.e., for an internally valid estimate). One such sufficient set of assumptions includes (1) on average, the outcomes of persons who received treatment a equal the potential outcomes of persons who received treatment a' had they received treatment a and vice versa (exchangeability), perhaps within strata of a set of covariates, Z (conditional exchangeability).[24] This is often referred to as the randomization or no unmeasured confounders assumption. (2) There is a nonzero probability of exposure within every stratum defined by Z (positivity, which holds trivially if Z is empty). (3) There are no versions of treatment other than those defined by A (treatment version irrelevance, sometimes referred to as consistency).[25–28] (4) One person's exposure does not affect another person's outcome (no interference).[29,30] (5) Outcome, treatment, and covariates are measured without error. We also require that all models be correctly specified, including the structural model and any parametric or semiparametric models used to describe associations between covariates and exposure or outcome. Most of these assumptions may be met in expectation if we conduct a randomized controlled trial. These assumptions may be less plausible in the observational setting where the treatment assignment mechanism is not known.[1,31,32]

Just as we never observe all potential outcomes for subjects in our study sample,[23,26,27] when we try to expand inference beyond the study sample to a particular target population, we typically do not observe any potential outcomes for subjects in our target population who were not selected into the study sample (unless additional data sources are available beyond the study sample). Assuming the sample average treatment effect is identifiable, the population average treatment effect will be identifiable when the sampling mechanism giving rise to the study sample is known (e.g., if the study sample is known to be a random sample from the target population). If the sampling mechanism is not known, additional assumptions are required to identify the population average treatment effect.

For external validity, it is sufficient to assume, first, that the participants included in the study sample are exchangeable with members of the target population who were not sampled, perhaps conditional on pretreatment characteristics W (conditional exchangeability between those sampled and those not sampled)[6]:

$$S \perp Y(x) \mid W \text{ for } x = a, a' \tag{3}$$

where S is an indicator of membership in the study sample. Enrollment into the sample is typically both under the control of the researcher (in designing a recruitment strategy) and under the control of the participants (in deciding whether to participate). The set of characteristics W should be chosen such that (3) is considered plausible. Judging whether a set of characteristics W is sufficient to satisfy this independence assumption may be a difficult task. One way to make this judgment more transparent is to explicitly represent the assumed data-generating mechanism using a directed acyclic graph (DAG).[33] The assumption encoded in (3) can then be verified by inspection of the DAG[2,6,34–36] as has been recommended for determining the set of covariates sufficient for confounder control for internal validity.[33,37,38] Second, we assume that, within strata of W, all subjects in the target population have some nonzero probability of being selected into the sample (analogous to positivity):

$$0 < P(S = 1 \mid W = w) \text{ for all } w \text{ such that } 0 < P(W = w)$$

Third, we assume the same distribution of versions of treatment in the study sample and the target population (treatment version irrelevance is a special case). This may be a strong assumption when the delivery mechanism for treatment differs dramatically between the study sample and the target population (e.g., treatment given to trial participants may have been accompanied by more adherence education and supportive services, as well as Hawthorne effects due to trial participation).[8,39] Fourth, we assume no interference[29,30] in the target population and the study sample (although these results can be extended to scenarios where the pattern of interference is the same in the target population and the study sample). Fifth, we assume no measurement error, including of W. We also require correct model(s) specification for any parametric or semiparametric models used to describe associations between covariates and outcome or any models used to describe the sampling mechanism. Assumptions sufficient for identification of a causal effect in the target population may, at first glance, look similar to those required for identification of a causal effect in the study sample.[5,40] However, assumptions about the relationships between the potential outcomes and the sampling mechanism are sufficient for external validity, compared with the case of internal validity for which assumptions about the relationships between the potential outcomes and the treatment assignment mechanism are sufficient. As assumptions sufficient for internal validity are met in expectation when treatment is randomized, assumptions sufficient for external validity will be met in expectation if the study sample is a simple random sample of the target population.

ESTIMATORS

If identifying assumptions hold, a generalization of the g-formula[13] (or adjustment formula)[14] or inverse probability of sampling weights can be employed to estimate the population average treatment effect. These estimators use data from the study sample on the exposure–outcome relationship and data from the target population on either (i) the distribution of W for the g-formula estimator or (ii) the sampling probabilities conditional on W for the inverse probability of sampling estimator.

Recall that the g-formula[13,14] to account for nonrandom treatment assignment is

$$E[Y(a)] = \sum_z E[Y|A = a, Z = z]P(Z = z)$$

where Z is a set of covariates sufficient for conditional exchangeability between treatment arms, that is, $A \perp Y(x) \mid Z$ for $x = a, a'$. Assuming the study sample is a random sample of the target population, nonparametric g-formula estimators will be consistent for the population average treatment effect. However, if the study sample was not randomly sampled from the target population, consistent estimators (based on the g-formula or otherwise) may not exist.[4]

Nonetheless, if in this setting we can find a set W that establishes conditional exchangeability between the sampled and unsampled, that is, (3), and treatment A is assigned at random to the study sample, then

$$E[Y(a)] = \sum_w E[Y \mid A = a, W = w, S = 1]P(W = w) \qquad (4)$$

A proof of this equivalence is given in Appendix A. Graphical conditions for determining the validity of (4) are provided in Bareinboim and Pearl.[3] If A was not randomly assigned in the sample and one is willing to assume for some set of covariates W' that (3) holds and also $A \perp Y(x) \mid W', S = 1$ for $x = a, a'$, then (4) holds with W' in place of W. In either case, the conditional expectation $E[Y \mid A = a, W = w, S = 1]$ is identifiable from the study sample. If an external source of data (not from the study sample) is available which identifies $P(W = w)$, then the population average treatment effect is identifiable via (4). This suggests the following substitution (or plug-in) estimator for the population average treatment effect:

$$\sum_w \left[\widehat{E}[Y|A = a, W = w, S = 1] - \widehat{E}[Y \mid A = a', W = w, S = 1]\right] \widehat{P}(W = w)$$

where $\widehat{E}[Y|A = x, W = w, S = 1]$ for $x = a, a'$ is based on data from the study sample and $\widehat{P}(W = w)$ is an estimator of the distribution of W in the target population based on external data.

The inverse probability of sampling weighted estimator arises from a different but equivalent expression for $E[Y(a)]$. In particular, again assuming conditional exchangeability between sampled and unsampled individuals (3), and random treatment assignment within the study sample, it follows that

$$E[Y(a)] = \frac{E[YI(A = a, S = 1)] / P(S = 1 \mid W)}{E[I(A = a, S = 1)] / P(S = 1 \mid W)}. \qquad (5)$$

A proof of this equivalence is also given in Appendix A. Expression (5) suggests instead the following plug-in estimator for the population average treatment effect:

$$\frac{\sum_{i=1}^{n} Y_i \mathrm{I}(A_i = a, S_i = 1)G(W_i)}{\sum_{i=1}^{n} \mathrm{I}(A_i = a, S_i = 1)G(W_i)} - \frac{\sum_{i=1}^{n} Y_i \mathrm{I}(A_i = a', S_i = 1)G(W_i)}{\sum_{i=1}^{n} \mathrm{I}(A_i = a', S_i = 1)G(W_i)}$$

where $G(w) = \hat{P}(S = 1 \mid W = w)^{-1}$ and $\hat{P}(S = 1 \mid W = w)$ is an estimator of the conditional probability of study enrollment based on an external source of data from the target population. If treatment is not randomly assigned within the study sample (e.g., as in an observational study), inverse probability of treatment weights[41] can be multiplied by $G(W_i)$ to simultaneously control for confounding.

The inverse probability of sampling and g-formula/transport formula estimators may give different results, particularly due to different modeling assumptions of the two approaches. A formal comparison of the two methods and a sufficient condition under which they will yield the same results when nonparametric estimators are employed is given in Appendix B. In settings where it is feasible to utilize both estimators, we will in general have a greater degree of confidence in the results when the two estimates are similar. Substantial differences between the two estimates could indicate possible violations of one or more of the assumptions being invoked.

EXAMPLE

To demonstrate how the methods described above can be used to estimate the population average treatment effect when the study sample is not a random sample of the target population, consider an arbitrarily large (infinite) target population where W_1 and W_2 are two independent Bernoulli random variables with expectations 0.15 and 0.20, respectively; $Y(1)$ and $Y(0)$ are Bernoulli random variables where $P(Y(a) = 1 \mid W_1, W_2) = 0.1073 - 0.05a + 0.20W_1 + 0.20W_2 - 0.15aW_1W_2$ for $a = 0, 1$; and A is Bernoulli with mean 0.5 and independent of $W_1, W_2, Y(1)$, and $Y(0)$. For this data-generating mechanism, $E[Y(1) = 1] = 0.123$ and $E[Y(0) = 1] = 0.177$, such that the population average treatment effect (×100%) is −5.5%. A study sample of $n = 2{,}000$ individuals was simulated by selecting a biased sample from the target population. Specifically, 320, 480, 480 and 720 individuals were randomly sampled from strata defined by $(W_1 = 0, W_2 = 0)$, $(W_1 = 1, W_2 = 0)$, $(W_1 = 0, W_2 = 1)$, and $(W_1 = 1, W_2 = 1)$. As in many trials, this sampling scheme oversampled participants at greater risk of the outcome ($W_1 = 1$ or $W_2 = 1$). For the $n = 2{,}000$ individuals in the study sample, the sample average treatment effect (×100%) was (0.250–0.356) × 100 = −10.7%. The simulated observed study data, W_1, W_2, A, and Y, are given in Table 1. Additionally, a random sample of $m = 50{,}000$ individuals from the target population was generated for which W_1, W_2, and S were observed (Table 2).

TABLE 1. Data from a Nonrandom Study Sample (n = 2,000)

W_1	W_2	A	Y	Number
0	0	0	0	143
0	0	0	1	22
0	0	1	0	141
0	0	1	1	14
0	1	0	0	172
0	1	0	1	72
0	1	1	0	178
0	1	1	1	58
1	0	0	0	166
1	0	0	1	82
1	0	1	0	171
1	0	1	1	61
1	1	0	0	162
1	1	0	1	207
1	1	1	0	248
1	1	1	1	103
Total:				2,000

W_1 and W_2 are pretreatment covariates that modify the effect of treatment A on outcome Y

TABLE 2. Data from a Random Sample from the Target Population (m = 50,000)

W_1	W_2	S	Number
0	0	0	33,498
0	0	1	320
0	1	0	8,162
0	1	1	480
1	0	0	5,556
1	0	1	480
1	1	0	784
1	1	1	720
Total:			50,000

W_1 and W_2 are the same pretreatment covariates defined in Table 1, measured in a random sample from the target population. S is an indicator of further selection from the target population into the study sample.

Given this realization of the data, the empirical risk difference (×100%) in the study sample between individuals with $A=1$ and $A=0$ is −13.1% (Table 3). On the other hand,[4] the estimate of the population average treatment effect (×100%) using the nonparametric g-(transport) formula is −5.4 and using inverse probability of sampling weights is −5.3%. As expected, the results from these two approaches are similar because we used nonparametric estimators of $P(Y=1 | A=a, W_1=w_1, W_2=w_2, S=1)$ and $P(W_1=w_1, W_2=w_2)$ in the g-formula approach and $P(S=1 | W_1=w_1, W_2=w_2)$ in the inverse probability of sampling approach. All calculations for this example appear in Appendix C.[29]

TABLE 3. Joint Distribution of Treatment A and Outcome Y in the Target Population and Data from Nonrandom Study Sample from Table 1 (n = 2,000)

Target Population (Probabilities)				Study Sample (n = 2,000)			
	$Y=0$	$Y=1$	Risk		$Y=0$	$Y=1$	Risk
$A=0$	0.414	0.089	0.177	$A=0$	643	383	0.373
$A=1$	0.439	0.061	0.123	$A=1$	738	236	0.242
Risk difference:		−0.055				−0.131	

TRANSPORTABILITY

Generalizing results to a target population which includes as members those persons included in the study sample differs from transporting results to a target population of which the study sample is not a subset.[42] That is, in a transportability problem,[43] the study sample is imagined to have arisen from a population that is distinct from the target population. Individuals within the target population have zero probability of being selected into the study sample when transporting study results, violating the positivity assumption as defined above. For transporting results, a different positivity assumption can be presumed

$$0 < P^*(S=1 | W=w) \text{for all } w \text{ such that } 0 < P(W=w)$$

where $P(\cdot)$ and $E(\cdot)$ denote probability and expectation in the target population as above, and $P^*(\cdot)$ and $E^*(\cdot)$ denote probability and expectation with respect to the superpopulation that gave rise to the study sample. Furthermore, the exchangeability assumption for transporting results is qualitatively different. When transporting results, it is sometimes sufficient to assume $E^*[Y(a) | S=1, W=w] = E[Y(a) | W=w]$ rather than (3). However, if S is associated with post-treatment covariates, there may not be any set of covariates W which satisfies $E^*[Y(a) | S=1, W=w] = E[Y(a) | W=w]$. Other distinctions between generalizability and transportability problems are beyond the scope of this article and are discussed elsewhere.[6] Methods for handling both selection and transportability problems are surveyed by Bareinboim and colleagues.[3,4,44]

DISCUSSION

To ensure an estimate is generalizable (in expectation) to a particular target population it would be sufficient to draw a study sample that is a random sample from that target population.[5] However, beyond the logistical, financial, and ethical challenges to conducting such a study, in certain circumstances, a study sample that is representative of the target population may be undesirable.[9,10] When first exploring the existence of a causal effect, epidemiologists may purposefully undertake nonrandom sample selection to increase statistical efficiency, to match or restrict on important confounders, or to allow estimation of subgroup effects.

Epidemiologists have been primarily concerned with the internal validity of effect estimates. However, the utility of an effect estimate for planning purposes and policy decision making will depend on both internal and external validity. For example, an internally valid estimate with extremely poor external validity may be of less use than an estimate with some internal bias but good external validity. External validity of an effect estimate will be threatened by the degree to which the prevalence of the effect measure modifiers differs in the study sample compared with the target population, as well as the magnitude of the modification.[7] For example, Greenhouse et al.[17] described a meta-analysis of trials of antidepressants in adolescents that suggested an increased risk of suicide among treated subjects. However, the majority of the meta-analyzed trials excluded participants with the most severe depression who would have experienced the greatest benefits from the therapy.[17] In this case, while trial effects were internally valid, the lack of external validity had serious implications for policy: the Food and Drug Administration used the meta-analysis to justify issuing a black box warning advising physicians and patients of increased suicide risk, which resulted in limiting potentially beneficial treatment options for depressed adolescent patients.[17,35] This example highlights the importance of balancing study design decisions to maximize both internal and external validity; internal and external bias both exist on a continuous scales (as degrees rather than as dichotomies) and relatively minor violations of internal validity may be tolerable in exchange for greater external validity.

Many of the assumptions and estimators we describe above may be familiar to the reader versed in threats to internal validity due to selection bias. Indeed, violations of Equation (3), in particular, may be interpreted as a selection bias problem (internal validity) or a generalizability problem (external validity). We view the process of enumerating the study sample as determining the generalizability or external validity of study results, while exclusion of participants due to drop out or missing data after the study sample has been defined determines the internal validity of study results. One might imagine the study as a randomized trial and ask whether selection occurred before or after enumeration of the study sample and treatment assignment (external or internal validity, respectfully). This distinction may be hypothetical, but is in harmony with existing thought experiments in epidemiology, such as framing analysis of observational data as if it arose from a randomized trial.[1,31,32]

We have discussed a g-formula estimator and inverse probability of sampling weighted estimator for generalizing results from a specified study sample to a specified target population. Doubly-robust estimation of the population average treatment effect is an area for future research. Such doubly-robust estimators would be consistent if either the model used to adjust for nonrandom sampling into the study or the model used to specify W -specific treatment effects is correct (without requiring both models be correct). Some doubly-robust estimators that might be easily adapted to the generalizability problem are those that have been developed for problems of missing outcome data in a trial.[45–47]

Commentaries on the lack of generalizability of randomized trials typically advocate evaluating a lengthy check list of potential determinants of external validity.[18,19,48,49] We argue that evaluations of generalizability could be more straightforward if considered quantitatively within the potential outcomes framework or the (logically equivalent) graphical models framework. Specifically, understanding the mechanism by which differences between the sample and the target populations arise is useful for identifying methods to account for those differences.

Finally, distinguishing internal and external threats to validity is useful for determining which parameters in the study sample or target population are estimable. When collider stratification bias due to selection is present in a study, it may threaten causal inference being made for any population,[50] even the study sample, and depending on the magnitude of the bias, may preclude attempts to generalize results to any specified target population. In contrast, if an analysis of a study is believed to have sufficient control of confounding and selection bias and differences in the average treatment effect can be attributed to nonrandom sampling of the study population, then (given the above assumptions) methods exist to generalize results to user-specified target population.[31] Understanding the source of the different biases that combine to influence a final estimate will help make analysis decisions that minimize the total bias. Generalizing effect estimates to the appropriate target population will improve their utility, and better inform implementation of interventions in target populations.

ACKNOWLEDGMENTS

The authors thank anonymous reviewers for their thoughtful comments that greatly improved this article.

REFERENCES

1. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60:578–586.
2. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference*. 2013;1:107–134.
3. Bareinboim E, Pearl J. Transportability of causal effects: completeness results. *AAAI Conference on Artificial Intelligence*. North America, 2012.
4. Bareinboim E, Tian J, Pearl J. Recovering from selection bias in causal and statistical inference. Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec City, Quebec, Canada. 2014:2410–2416.
5. 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*. 2001;174:369–386.
6. Pearl J. Generalizing experimental findings. *J Causal Inference*. 2015;3:259–266.
7. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am J Epidemiol*. 2010;172:107–115.

8. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011;22:368–377.
9. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol*. 2013;42:1012–1014.
10. Rothman K, Hatch E, Gallacher J. Representativeness is not helpful in studying heterogeneity of effects across subgroups. *Int J Epidemiol*. 2014;43:633–634.
11. Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. *J R Stat Soc Ser A Stat Soc*. 2016;179:319–376.
12. Hotz VJ, Imbens GW, Mortimer JH. Predicting the efficacy of future training programs using past experiences at other locations. *J Econom*. 2005;125:241–270.
13. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393–1512.
14. Pearl J. *Causality*. New York, NY: Cambridge University Press; 2009.
15. Maldonado G, Greenland S. Estimating causal effects. *Int J Epidemiol*. 2002;31:422–429.
16. Gandhi M, Ameli N, Bacchetti P, et al. Eligibility criteria for HIV clinical trials and generalizability of results: the gap between published reports and study protocols. *AIDS*. 2005;19:1885–1896.
17. 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;27:1801–1813.
18. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365:82–93.
19. Rothwell PM. Commentary: external validity of results of randomized trials: disentangling a complex concept. *Int J Epidemiol*. 2010;39:94–96.
20. Heckman JJ. Sample selection bias as a specification error. *Econometrica*. 1979;47:153–161.
21. Angrist JD. Conditional independence in sample selection models. *Econom Lett*. 1997;54:103–112.
22. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:945–960.
23. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol*. 2015;44:1452–1459.
24. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011;67:1406–1413.
25. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20:3–5.
26. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21:872–875.
27. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20:880–883.
28. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008;32(Suppl 3):S8–S14.
29. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc*. 2008;103:832–842.
30. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. *Stat Methods Med Res*. 2012;21:55–75.
31. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
32. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183:758–764.
33. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
34. Petersen ML. Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology*. 2011;22:378–381.
35. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21:243–256.
36. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci*. 2014;29:579–595.
37. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
38. Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR. "Toward a clearer definition of confounding" revisited with directed acyclic graphs. *Am J Epidemiol*. 2012;176:506–511.
39. Westreich D, Edwards JK. Invited commentary: every good randomization deserves observation. *Am J Epidemiol*. 2015;182:857–860.
40. Tipton E. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J Educ Behav Stat*. 2013;38:239–266.
41. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
42. Hoggatt KJ, Greenland S. Commentary: extending organizational schema for causal effects. *Epidemiology*. 2014;25:98–102.
43. Pearl J, Bareinboim E. External validity and transportability: a formal approach. *Joint Statistical Meetings*. Miami Beach, FL. 2011;157–171.
44. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 2016;113:7345–7352.
45. Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat Sci*. 2007;22:544–559.
46. Colantuoni E, Rosenblum M. Leveraging prognostic baseline variables to gain precision in randomized trials. *Stat Med*. 2015;34:2602–2617.
47. Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med*. 2009;28:39–64.
48. Szklo M. Population-based cohort studies. *Epidemiol Rev*. 1998;20:81–90.
49. 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;39:89–94.
50. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.

APPENDIX A

Proof of (4):

$$E[Y(a)]=\sum_w E[Y(a)\mid W=w]P(W=w) \quad \text{Law of total probability}$$

$$=\sum_w E[Y(a)\mid W=w,S=1]P(W=w) \quad S\perp Y(a)\mid W$$

(conditional exchangeability between sampled and nonsampled)

$$=\sum_W E[Y(a)\mid A=a,W=w,S=1]P(W=w) \quad A\perp Y(a)\mid W,S=1$$

(conditional exchangeability between treated and untreated)

$$=\sum_w E[Y\mid A=a,W=w,S=1]P(W=w)$$

(consistency/treatment version irrelevance)

Proof of (5):

Similar to the proof of (4), under the stated assumptions it is straightforward to show:

$$E\left[\frac{YI(A=a,S=1)}{P(S=1|W)}\right]=E[Y(a)]P(A=a|S=1)$$

and

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

which together imply (5).

APPENDIX B

Nonequivalence of the nonparametric g-formula and inverse probability of sampling weighted estimator when weights are estimated nonparametrically.

Suppose we observe (W_i, A_i, Y_i) for $i = 1,...,n$ individuals in the study sample. Suppose we also observe (W_j, S_j) for $j = 1,...,m$ individuals based on a random sample from the target population.

The nonparametric g-formula based estimator is as follows:

$$E[\widehat{Y(a)}] = \sum_w \hat{E}(Y|A = a, W = w, S = 1)\hat{P}(W = w)$$

where the conditional expectation is estimated by data from the study sample, that is,

$$\hat{E}(Y|A = a, W = w, S = 1) = \frac{\sum_{i=1}^{n} Y_i I[A_i = a, W_i = w]}{\sum_{i=1}^{n} I[A_i = a, W_i = w]}$$

and $\hat{P}(W = w)$ is a nonparametric estimator based on data from the target population, that is,

$$\hat{P}(W = w) = \sum_{j=1}^{m} I(W_j = w)/m$$

For notational convenience, let
$f(w) = \sum_{i=1}^{n} Y_i I[A_i = a, W_i = w]$ and
$g(w) = \sum_{i=1}^{n} I[A_i = a, W_i = w]$ so that the nonparametric g-formula estimator can be written

$$E[\widehat{Y(a)}] = \sum_w \frac{f(w)}{g(w)}\hat{P}(W = w) \tag{6}$$

Now consider the nonparametric inverse probability of sampling weighted estimator is

$$E[\widehat{Y(a)}] = \frac{\sum_{i=1}^{n} Y_i I[A_i = a, S_i = 1]G_i(W_i)}{\sum_{i=1}^{n} I[A_i = a, S_i = 1]G_i(W_i)}$$

where $G_i(W_i)$ is the inverse of the estimated probability of being sampled conditional on covariates W_i. Equivalently, we can write the inverse probability of sampling weighted estimator as

$$E\left[\widehat{Y(a)}\right] = \frac{\sum_{i=1}^{n}\sum_w Y_i I(W_i = w, A_i = a)G(w)}{\sum_{i=1}^{n}\sum_w I(W_i = w, A_i = a)G(w)}$$

or

$$E\left[\widehat{Y(a)}\right] = \frac{\sum_w\left[G(w)\sum_{i=1}^{n} Y_i I(W_i = w, A_i = a)\right]}{\sum_w\left[G(w)\sum_{i=1}^{n} I(W_i = w, A_i = a)\right]}$$

Using the notation above, we have

$$E\left[\widehat{Y(a)}\right] = \frac{\sum_w G(w)f(w)}{\sum_w G(w)g(w)}$$

Assume we estimate $P(S = 1|W)$ nonparametrically such that

$$G(w) = \frac{\sum_{j=1}^{m} I(W_j = w)}{\sum_{j=1}^{m} I(W_j = w, S_j = 1)} = \frac{\hat{P}(W = w)}{\hat{P}(W = w, S = 1)}$$

where $\hat{P}(W = w, S = 1) = \sum_{j=1}^{m} I\left(W_j = w, S_j = 1\right)/m$. Then

$$E\left[\widehat{Y(a)}\right] = \frac{\sum_w f(w)\hat{P}(W = w)/\hat{P}(W = w, S = 1)}{\sum_w g(w)\hat{P}(W = w)/\hat{P}(W = w, S = 1)} \tag{7}$$

Thus the nonparametric g-formula estimator (6) and the nonparametric inverse probability of sampling weighted estimator (7) will be equal if

$$g(w) = \hat{P}(W = w, S = 1)\sum_v g(v)\hat{P}(W = v)/\hat{P}(W = v, S = 1)$$

that is,

$$\frac{g(w)}{\hat{P}(W = w, S = 1)} = \sum_v \frac{g(v)}{\hat{P}(W = v, S = 1)}\hat{P}(W = v)$$

But this need not be true in general.

APPENDIX C

Calculations for example

Population average treatment effect (*PATE*):

$$E[Y(1) - Y(0)] = 0.1228 - 0.1773 = -0.0545$$

Sample average treatment effect (*SATE*):

$$0.2495 - 0.3560 = -0.1065$$

Empirical risk difference in the study sample:

$$\sum_{i=1}^{n}\left[\frac{Y_i I[A_i = 1]}{I[A_i = 1]} - \frac{Y_i I[A_i = 0]}{[A_i = 0]}\right] = \frac{236}{974} - \frac{383}{1026} = -0.1310$$

Estimation of the *PATE* from the study sample using the g-formula:

$$\sum_w\left[\frac{\sum_{i=1}^{n} Y_i I[A_i = 1, W_i = w]}{\sum_{i=1}^{n} I[A_i = 1, W_i = w]}\right]\left[\frac{\sum_{j=1}^{m} I(W_j = w)}{m}\right] - \sum_w\left[\frac{\sum_{i=1}^{n} Y_i I[A_i = 0, W_i = w]}{\sum_{i=1}^{n} I[A_i = 0, W_i = w]}\right]\left[\frac{\sum_{j=1}^{m} I(W_j = w)}{m}\right]$$

$$= \left[\frac{14}{155}\times\frac{33818}{50000} + \frac{58}{236}\times\frac{8642}{50000} + \frac{61}{232}\times\frac{6036}{50000} + \frac{103}{351}\times\frac{1504}{50000}\right] - \left[\frac{22}{165}\times\frac{33818}{50000} + \frac{72}{244}\times\frac{8642}{50000} + \frac{82}{248}\times\frac{6036}{50000} + \frac{207}{369}\times\frac{1504}{50000}\right]$$

$$= 0.1441 - 0.1980 = -0.0538$$

Estimation of the *PATE* from the study sample using inverse probability weighting:

First, estimate $G(w) = \left[\widehat{P}(S = 1 \mid W = w)\right]^{-1}$.

$$\left[\widehat{P}(S = 1 \mid W_1 = 0, W_2 = 0)\right]^{-1} = 33818/320 = 105.68$$

$$\left[\widehat{P}(S = 1 \mid W_1 = 0, W_2 = 1)\right]^{-1} = 8642/480 = 18.00$$

$$\left[\widehat{P}(S = 1 \mid W_1 = 1, W_2 = 0)\right]^{-1} = 6036/480 = 12.58$$

$$\left[\widehat{P}(S = 1 \mid W_1 = 1, W_2 = 1)\right]^{-1} = 1504/720 = 2.09$$

Then the inverse probability of sampling weighted estimate equals:

$$\frac{\sum_w \left[G(w)\sum_{i=1}^{n} Y_i I(W_i = w, A_i = 1)\right]}{\sum_w \left[G(w)\sum_{i=1}^{n} I(W_i = w, A_i = 1)\right]} - \frac{\sum_w \left[G(w)\sum_{i=1}^{n} Y_i I(W_i = w, A_i = 0)\right]}{\sum_w \left[G(w)\sum_{i=1}^{n} I(W_i = w, A_i = 0)\right]}$$

$$= \frac{105.68\times 14 + 18.00\times 58 + 12.58\times 61 + 2.09\times 103}{105.68\times 155 + 18.00\times 236 + 12.58\times 232 + 2.09\times 351} - \frac{105.68\times 22 + 18.00\times 72 + 12.58\times 82 + 2.09\times 207}{105.68\times 165 + 18.00\times 244 + 12.58\times 248 + 2.09\times 369}$$

$$= \frac{3506.01}{24280.18} - \frac{5084.84}{25719.82} = 0.1444 - 0.1977 = -0.0533$$