Forthcoming, Journal of Causal Inference,
Causal, Casual, and Curious Section.

TECHNICAL REPORT
R-452
July 2015

Generalizing Experimental Findings

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
(310) 825-3243 / judea@cs.ucla.edu

Abstract

This note examines one of the most crucial questions in causal inference: "How generalizable are randomized clinical trials?" The question has received a formal treatment recently, using a non-parametric setting, and has led to a simple and general solution. I will describe this solution and several of its ramifications, and compare it to the way researchers have attempted to tackle the problem using the language of ignorability. We will see that ignorability-type assumptions need to be enriched with structural assumptions in order to capture the full spectrum of conditions that permit generalizations, and in order to judge their plausibility in specific applications.

Keywords: Generalizability, transportability, selection bias, admissibility, ignorability

1 Transportability and Selection Bias

The long-standing problem of generalizing experimental findings from the trial sample to the population as a whole, also known as the problem of "sample selection-bias" (Heckman, 1979; Bareinboim et al., 2014), has received renewed attention in the past decade, as more researchers come to recognize this bias as a major threat to the validity of experimental findings in both the health sciences (Stuart et al., 2015) and social policy making (Manski, 2013).

Since participation in a randomized trial cannot be mandated, we cannot guarantee that the study population would be the same as the population of interest. For example, the study population may consist of volunteers, who respond to financial and medical incentives offered by pharmaceutical firms or experimental teams, so, the distribution of outcomes in the study may differ substantially from the distribution of outcomes under the policy of interest.

Another impediment to the validity of experimental finding is that the types of individuals in the target population may change over time (Hotz et al., 2005). For example, as more individuals become eligible for health insurance, the types of individuals seeking services would no longer match the type of individuals that were sampled for the study (Stuart et al., 2015). A similar change would occur as more individuals become aware of the efficacy

of the treatment. The result is an inherent disparity between the target population and the population under study.

The problem of generalizing across disparate populations has received a formal treatment in (Pearl and Bareinboim, 2014) where it was labeled "transportability," and where necessary and sufficient conditions for valid generalization were established (see also Bareinboim and Pearl, 2013). The problem of selection bias, though it has some unique features, can also be viewed as a nuance of the transportability problem, thus inheriting all the theoretical results established in (Pearl and Bareinboim, 2014) that guarantee valid generalizations. I will describe the two problems side by side and then return to the distinction between the type of assumptions that are needed for enabling generalizations.

The transportability problem concerns two dissimilar populations, Π and Π^*, and requires us to estimate the average causal effect $P^*(y_x)$ (explicitly: $P^*(y_x) \stackrel{\Delta}{=} P^*(Y = y|do(X = x))$ in the target population Π^*, based on experimental studies conducted on the source population Π.[1] Formally, we assume that all differences between Π and Π^* can be attributed to a set of factors S that produce disparities between the two, so that $P^*(y_x) = P(y_x|S = 1)$. The information available to us consists of two parts; first, treatment effects estimated from experimental studies in Π and, second, observational information extracted from both Π and Π^*. The former can be written $P(y|do(x), z)$, where Z is set of covariates measured in the experimental study, and the latters are written $P^*(x, y, z) = P(x, y, z|S = 1)$, and $P(x, y, z)$ respectively. In addition to this information, we are also equipped with a qualitative causal model M, that encodes causal relationships in Π and Π^*, with the help of which we need to identify the query $P^*(y_x)$. Mathematically, identification amounts to transforming the query expression

$$P^*(y_x) = P(y|do(x), S = 1) \tag{1}$$

into a form derivable from the available information I_{TR}, where

$$I_{TR} = \{P(y|do(x), z), P(x, y, z), P(x, y, z|S = 1)\}. \tag{2}$$

The first two components of I_{TR} represent, respectively, the experimental and observational findings in Π, while the third component represents observational findings in Π^*. Appendix 1 demonstrates how the query $P^*(y_x)$ can be derived from I_{TR} using assumptions about the disparities between Π and Π^* that are encoded in a graph.

The selection bias problem is slightly different. Here the aim is to estimate the average causal effect $P(y_x)$ in the Π population, while the experimental information available to us, I_{SB}, comes from a preferentially selected sample, $S = 1$, and is given by $P(y|do(x), z, S = 1)$. In addition, we also assume to have access to observational information $P(x, y, z|S = 1)$ and $P(x, y, z)$; the first represents observations obtained from the selected sample, $S = 1$, and the second represents observation taken on the population at large. Thus, the selection bias problem calls for transforming the query $P(y_x)$ to a form derivable from the information set:

$$I_{SB} = \{P(y|do(x), z, S = 1), P(x, y, z|S = 1), P(x, y, z)\}. \tag{3}$$

[1] We focus our discussion on the average causal effect (ATE), yet identical considerations apply to other causal parameters, such as the effect of treatment on the treated (ETT). On the connection between ATE and ETT, see (Shpitser and Pearl, 2009).

In the Appendix section, we demonstrate how transportability problems and selection bias problems are solved using the transformations described above. At this point, however, it is important to note the syntactic differences between the information sets available in the two problems. I_{TR} is characterized by the fact that S does not appear in the conditioning part of any *do*-expression, thus reflecting the fact that we do not have experimental information from the target population Π^*. I_{SB} on the other hand is characterized by the fact that *do*-expressions are always conditioned on S, reflecting the fact that we have experimental information only on the selected sample, $S = 1$.

The analysis reported in (Pearl and Bareinboim, 2014) has resulted in an algorithmic criterion for deciding whether transportability is feasible and, when confirmed, the algorithm produces an estimand for the desired effects (Bareinboim and Pearl, 2013). The algorithm is complete, in the sense that, when it fails, a consistent estimate of the target effect does not exist (unless one strengthens the assumptions encoded in M).

There are several lessons to be learned from this analysis when considering generalizing experimental findings.

1. The graphical criteria that authorize transportability are applicable to selection bias problems as well, provided that the graph structures for the two problems are identical. This means that whenever a selection bias problem is characterized by a graph for which transportability is feasible, recovery from selection bias is feasible by the same algorithm. (The Appendix demonstrates this correspondence.)

2. The assumptions needed for transportability are more involved than the ones usually invoked for ensuring non-confoundedness, also called "treatment assignment ignorability." In graphical terms, these assumptions may require several d-separation tests on several sub-graphs. It is utterly unimaginable therefore that such assumptions could be managed by unaided human judgment, as is normally assumed in the potential outcomes literature (Hartman et al., 2015; Stuart et al., 2015).

3. In general, problems associated with generalizing across populations cannot be handled by balancing disparities between distributions. A given disparity between $P(x, y, z)$ and $P^*(x, y, z)$ may demand different adjustments, depending on the location of S in the causal structure. A simple example of this phenomenon is demonstrated in Fig. 3(b) of (Pearl and Bareinboim, 2014) where a disparity in the average reading ability of two cities requires two different treatments, depending on what causes the disparity. If the disparity emanates from age differences, adjustment is necessary, because age is likely to affect the potential outcomes. If, on the other hand the disparity emanates from differences in educational programs, no adjustment is needed, since education, in itself, does not modify response to treatment. Such distinctions, which may become quite intricate in large systems, are managed automatically in the graph-based representation.

4. In many instances, generalizations can only be achieved by conditioning on post-treatment variables, an operation that is generally frowned upon in the potential outcomes framework (Rosenbaum, 2002, pp. 73–74; Rubin, 2004; Sekhon, 2009) but has

become extremely useful in graphical analysis. The difference between the conditioning operators used in these two frameworks is reflected in the difference between the counterfactual expression $P(Y_x = y|z)$ and the *do*-expression $P(Y = y|do(X = x), z)$. (Pearl, 2015). The latter expression defines information that is estimable directly from experimental studies, whereas the former invokes retrospective counterfactuals that may or may not be estimable empirically.

In the next Section we will discuss the differences between these two conditioning operators and the benefit of leveraging post-treatment variables in problems concerning generalization.

2 Ignorability versus Admissibility in the Pursuit of Generalizations

A key assumption in almost all conventional analyses of generalization (from sample-to-population) is S-ignorability, written

$$Y_x \perp\!\!\!\perp S|Z \tag{4}$$

where Y_x is the potential outcome predicated on the intervention $X = x$, S is a selection indicator (with $S = 1$ standing for selection into the sample) and Z a set of observed covariates. This assumption, commonly written as a difference $Y_1 - Y_0 \perp\!\!\!\perp S|Z$, appears in Hotz et al. (2005); Cole and Stuart (2010); Tipton et al. (2014); Hartman et al. (2015), and possibly other researchers confined to potential outcomes analysis. This assumption states that in every stratum $Z = z$ of the set Z, the potential outcome Y_x is independent of the factors S that may produce cross-population differences.

Given this assumption, the problem of generalizing across populations has a trivial solution, which reads: If we succeed in finding a set Z of pre-treatment covariates such that cross-population differences disappear in every stratum $Z = z$, then the problem can be solved by averaging over those strata.[2]

Specifically, if $P(y_x|S = 1, Z = z)$ is the z-specific probability distribution of Y_x in the sample, then the distribution of Y_x in the population at large is given by the *post-stratification* formula

$$P(y_x) = \sum_z P(y_x|S = 1, z)P(z) \tag{5}$$

which is often referred to as *re-calibration* or *re-weighting*. Here, $P(z)$ is the probability of $Z = z$ in the target population (where $S = 0$). Equation (5) follows from S-ignorability by conditioning on z and, adding $S = 1$ to the conditioning set – a one-line proof. The proof fails however when no covariate set Z exists that satisfies S-ignorability, in which case the post-stratification formula will be invalid. Moreover, even when S-ignorability holds, Eq. (5) would only be applicable if the factor $P(y_x|S = 1, z)$ is estimable in the experimental study

[2]Lacking a procedure for finding Z, this solution addresses only part of the problem, leaving the choice of Z to unaided intuitive judgement.

and this will generally not be the case when Z contains post-treatment variables (see Pearl 2015, Fig. 1).

Symmetrically, when we consider transportability problems, our query is $P^*(y_x) = P(y|do(x), S = 1)$ (see Eq. (1)), and S-ignorability would permit us to remove the $S = 1$ condition and obtain the post-stratification formula

$$P^*(y_x) = P(y_x|S = 1) = \sum_z P(y_x|z)P(z|S = 1) \tag{6}$$

Similar to Eq. (5), this formula takes a weighted average of the z-specific potential outcome Y_x over all levels of Z. Here, in syntactic contrast, the average is weighed by $P(z|S = 1)$ which is, again, the distribution of Z in the target population (where $S = 1$). As in the case of selection bias, Eq. (6) is only useful when S-ignorability holds and when $P(y_x|z)$ is estimable from the experimental data. Unfortunately, when Z contains post-treatment variables, the former condition will be harder to meet; we shall see that S-ignorability is rarely satisfied in transportability problems by any set Z containing post-treatment variables.

In graphical analysis, on the other hand, the problem of generalization has been studied using another assumption, labeled S-admissibility (Pearl and Bareinboim, 2014), which is defined by:

$$P(y|do(x), z) = P(y|do(x), z, s) \tag{7}$$

or, using counterfactual notation,

$$P(y_x|z_x) = P(y_x|z_x, s_x)$$

It states that in every treatment regime $X = x$, the observed outcome Y is conditionally independent of the selection mechanism S, given Z, all evaluated at that same treatment regime.

Clearly, S-admissibility coincides with S-ignorability for pretreatment S and Z; the two notions differ however for treatment-dependent selection and covariates. To witness, consider the model of Fig. 1(a), and let X stand for education, Z for skill, S for training, and Y for salary. S-admissibility (4) looks at those people who were assigned x years of education who subsequently achieved skill level z, and asks whether their salary Y would depend on their training S. The graph states that skill alone determines salary, not how it was acquired, therefore $P(y|do(x), z)) = P(y|do(x), z, s) = P(y|z)$ namely, training and education have no effect on salary, once we know z, as shown in the graph.

In contrast, S-ignorability $Y_x \perp\!\!\!\perp S|Z$ asks for the role that training plays in the salary of those individuals who are currently at skill $Z = z$, had they received x years of schooling. Surely, unless x is pathologically low, the skill levels attained by these individuals would depend on the amount of training (S) they receive, and so would their salary Y. We thus conclude that Y_x is not independent of S given Z, namely, S-ignorability does not hold. The condition $Z = z$ merely selects a subpopulation for consideration but, unless individuals in this subpopulation possess some abnormal qualities, they should exhibit the natural dependence of salary on training.[3]

[3]To show explicitly that S-ignorability does not hold in Fig. 1(a), one can examine a linear model and

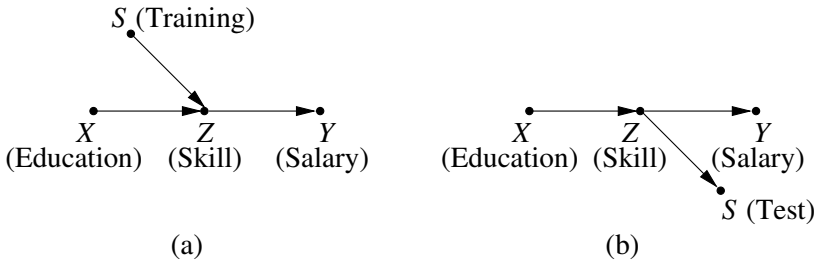

Figure 1: (a) A transportability model in which a post-treatment variable Z is S-admissible but not S-ignorable; (b) A selection-bias model in which Z is both S-admissible and S-ignorable. Note that S is a root node in (a) and a sink node in (b), where it is a proxy of Z. In both models, the post-stratification formula (5) is not estimable non-parametrically.

The Appendix section shows that unbiased generalization across studies is indeed feasible in scenarios like Fig. 1 (a), despite the fact that Z is not S-ignorable. This is facilitated by the fact that Z is S-admissible, since Z separates Y from S in the graph, and leads to the following estimand for the target effect:

$$P(y_x|S=1) = \sum_z P(y|do(x),z)P(z|x,S=1).$$

Note that this estimand invokes nonconventional average of the z-specific effect, weighted by the conditional probability $P(z|x)$ at the target population.

A similar situation occurs in sample-selection problems such as the one depicted in Fig. 1(b), where generalization from samples to populations through the post-stratification formula (5) requires S-ignorability. Here, the post-stratification formula (5) is valid because Z is S-ignorable (Z separates S from Y_x in the graph), yet the formula is useles, because the z-specific causal effect $P(y_x|S=1,z)$ is not estimable from the experimental study.

Remarkably, the target distribution $P(y_x)$ can be estimated using a modified formula:

$$P(y_x) = \sum_z P(y|do(x),z,S=1)P(z|x)$$

which follows from the fact that Z is S-admissible. The derivation is presented in Scenario 3 of the Appendix and demonstrates that, regardless of whether Z satisfies S-ignorability or S-admissibility, experimental findings are not generalizable by standard procedures of post-stratification. Rather, modified procedures need be applied, dictated by the graph structure.

One of the reasons that S-admissibility has received greater attention in the graph-based literature is that it has a very simple graphical representation: Z and X should separate Y from S in a mutilated graph, from which all arrows entering X have been removed. Such a

use Eq. (11.28) of (Pearl, 2009, p. 389) and show that it yields

$$E[Y_x|Z=z,S=s] = ax + bz + cs$$

with non-zero c.

graph depicts conditional independencies among observed variables in the population under experimental conditions, i.e., where X is randomized.

S-ignorability requires a more elaborate graphical interpretation; it can be verified from either twin networks (Pearl, 2009, pp. 213-4) or from counterfactually augmented graphs (Pearl, 2009, p. 341). Using either representation, it is easy to see that S-ignorability is rarely satisfied in problems in which Z is a post-treatment variable. This is because, whenever S is an ancestor of Z, or a proxy of such ancestor, Z cannot separate Y_x from S.

As noted in (Keiding, 1987) the re-calibration formula (5) goes back to 18th century demographers (Dale, 1777; Tetens, 1786) facing the task of predicting overall mortality (across populations) from age-specific data. Their reasoning was probably as follows: If the source and target populations differ in distribution by a set of attributes Z, then to correct for these differences we need to weight samples by a factor that would restore similarity to the two distributions. Some researchers view Eq. (5) as a version of Horvitz and Thompson (1952) post-stratification method of estimating the mean of a super-population from un-representative stratified samples. The essential difference between survey sampling calibration and the calibration required in Eq. (5) is that the calibrating covariates Z are not just any set by which the distributions differ; they must satisfy the S-ignorability (or admissibility) condition, which is a causal, not a statistical condition and is not discernible therefore from distributions over observed variables. In other words, the re-calibration formula should depend on disparities between the causal models of the two populations, not merely on distributional disparities; we discussed this point in Section 1 (item 3) and it is also demonstrated in the Appendix (Fig. 2(a)).

While S-ignorability and S-admissibility are both sufficient for re-calibrating pre-treatment covariates Z, S-admissibility goes further and discovers generalizations that leverage both pre-treatment and post-treatment variables. The three examples discussed in the Appendix demonstrate this point.

Conclusions

1. Many opportunities for generalization are opened up through the use of post-treatment variables. These opportunities remain inaccessible to ignorability-based analysis, partly because S-ignorability does not always hold for such variables but, mainly, because ignorability analysis requires information in the form of z-specific counterfactuals, which is often not estimable from experimental studies.

2. Most of these opportunities have been chartered through the completeness results for transportability (Heckman, 1979), others can be revealed by simple derivations in *do*-calculus as shown in the Appendix.

3. There is still the issue of assisting researchers in judging whether S-ignorability (or S-admissibility) is plausible in any given application. Graphs excel in this dimension because they match the format in which people store scientific knowledge. Researchers who insist on discerning S-ignorability by appealing to human intuition do so at the peril of missing opportunities for generalization, or producing biased effect estimates. Readers can appreciate the magnitude of these perils by examining the simple examples

presented in Fig. 2 of the Appendix; discerning S-ignorability in any one of the three scenarios is a formidable judgmental task if unaided by graphs.

Acknowledgment

This note has benefitted from discussions with Elias Bareinboim, Stephen Cole, Peng Ding, Guido Imbens, Jasjeet Sekhon, and Elizabeth Tipton.

This research was supported in parts by grants from NSF #IIS-1302448 and ONR #N00014-10-1-0933 and #N00014-13-1-0153.

Appendix

To each of the models represented in Fig. 2 we will provide a scenario, a problem specification and a derivation of the target estimand.

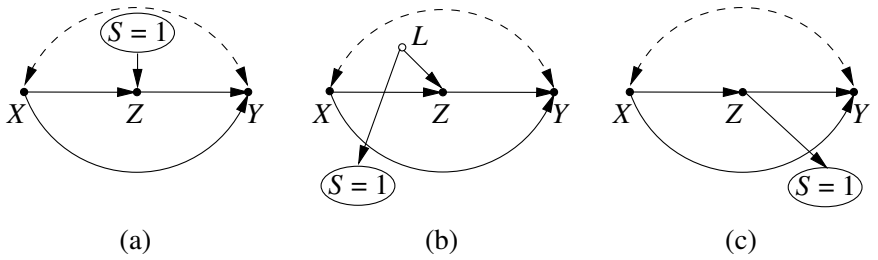

Figure 2: (a) Generalizable transportability problem in which Z is S-admissible but S-ignorability does not hold. (b) Generalizable selection-bias problem in which Z is S-admissible but S-ignorability does not hold. (c) Generalizable selection-bias problem in which S-admissibility and S-ignorability both hold, yet post-stratification (Eq. (5)) fails to estimate the target treatment effect $P(y_x)$.

Scenario 1 *(Figure 2(a)):*
X = Treatment, Y = outcome, Z = a bio-marker believed to mediate between treatment and outcome. S = a factor (say diet) that makes the effect of X on Z different in the two populations, Π and Π^*. The curved dashed arch between X and Y represents the presence of unobserved confounders.

Problem formulation:
Needed:

$$P^*(y_x) = P(y|do(x), S = 1)$$

Information set available:

$$I_{TR} = \{P(y|do(x), z), P(x, y, z|S = 1), P(x, y, z)\}.$$

Assumptions: S-admissibility (deduced from Fig. 2(a))

$$P(y|do(x), z) = P(y|do(x), z, s)$$

Derivation:

$$\begin{aligned} P^*(y_x) &= P(y|do(x), S=1) \\ &= \sum_z P(y|do(x), S=1, z) P(z|do(x), S=1) \\ &= \sum_z P(y|do(x), z) P(z|do(x), S=1) \\ &= \sum_z P(y|do(x), z) P(z|x, S=1) \end{aligned}$$

Each step in this derivation follows from probability theory and the assumption of S-admissibility which permits us to remove the factor $S=1$ from the first factor of the second line. The result is an estimand in which the condition $S=1$ does not appear in any *do*-expression, hence it is estimable from I_{TR}.

Scenario 2 *(Figure 2(b))*
This is a selection-bias version of the transportability problem presented in Scenario 1. Assume variable L stands for "location" and that selection for the study prefers subjects from one location over another (Hotz et al., 2005). The task is to estimate the average causal effect over the entire population.

Problem formulation:
Needed:

$$P(y_x) = P(y|do(x))$$

Information set available:

$$I_{SB} = \{P(y|do(x), z, S=1), P(x, y, z|S=1), P(x, y, z)\}.$$

Assumptions: S-admissibility (deduced from the model of Fig. 2(b))

$$P(y|do(x), z) = P(y|do(x), z, s)$$

Derivation:

$$\begin{aligned} P(y_x) &= P(y|do(x)) \\ &= \sum_z P(y|do(x), z) P(z|do(x)) \\ &= \sum_z P(y|do(x), z, S=1) P(z|do(x)) \\ &= \sum_z P(y|do(x), z, S=1) P(z|x) \end{aligned}$$

The first term in the sum is estimable from the biased experimental study while the second from the target population.

Scenario 3 *(Figure 2(c))*
This is another selection-bias version of the problem presented in Scenario 1. Assume Z represents a post-treatment complication and, naturally, people with complications are more likely to enter the database.

Problem formulation:

The problem is identical to that of Scenario 2 with the exception that now both S-admissibility and S-ignorability hold for variable Z. The former can be seen from its graphical definition, since Z and X separate Y from S, and the latter by noting the Z separate S from all exogenous factors that affect Y.
Derivation:
The same as in Scenario 2. Again, we see that the final estimand calls for averaging the z-specific effect in the experiment over all strata of Z, but now the average is weighted by the conditional probability $P(z|x)$ instead of the marginal $P(z)$ that appears in Eq. (5).

Remark 1 *Note that, in Scenario 2, if variable L is observable, then the selection bias problem can be solved by re-calibration over L, since L is treatment-independent and satisfies S-ignorability (and S-admissibility). It is only when L is unobserved that we must resort to Z, a post treatment variable that does not satisfy S-ignorability.*

References

Bareinboim, E. and Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* **1** 107–134.

Bareinboim, E., Tian, J. and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* (C. E. Brodley and P. Stone, eds.). AAAI Press, Palo Alto, CA. Best Paper Award, <http://ftp.cs.ucla.edu/pub/stat_ser/r425.pdf>.

Cole, S. and Stuart, E. (2010). Generalizing evidence from randomized clinical trials to target populations. *American Journal of Epidemiology* **172** 107–115.

Dale, W. (1777). A Supplement to Calculations of the Value of Annuities, Published for the Use of Societies Instituted for Benefit of Age Containing Various Illustration of the Doctrine of Annuities, and Compleat Tables of the Value of 1£. Immediate Annuity.

Hartman, E., Grieve, R., Ramsahai, R. and Sekhon, J. (2015). From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects. *Journal Royal Statistical Society: Series A (Statistics in Society)* **178** 757–778.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

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

Hotz, V. J., Imbens, G. W. and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* **125** 241–270.

Keiding, N. (1987). The method of expected number of deaths, 1786–1886–1986, correspondent paper. *International Statistical Review* **55** 1–20.

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press, Cambridge, MA.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2015). Conditioning on post-treatment variables. *Journal of Causal Inference* **3** 131–137.

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

Rosenbaum, P. (2002). *Observational Studies*. 2nd ed. Springer-Verlag, New York.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *The Annual Review of Political Science* **12** 487–508.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (J. Bilmes and A. Ng, eds.). AUAI Press, Montreal, Quebec, 514–521.

Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science* **16** 475–485.

Tetens, J. (1786). *Einleitung zur Berechnung der Leibrenten und Anwartschaften II*. Weidmanns Erben und Reich, Leipzig.

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