Forthcoming, Sociological Methods and Research.

TECHNICAL REPORT
R-406
First version December 2012
Revised July 2015

Detecting Latent Heterogeneity

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

Abstract

We address the task of determining, from statistical averages alone, whether a population under study consists of several subpopulations, unknown to the investigator, each responding to a given treatment markedly differently. We show that such determination is feasible in three cases: (1) randomized trials with binary treatments, (2) models where treatment effects can be identified by adjustment for covariates, and (3) models in which treatment effects can be identified by mediating instruments. In each of these cases we provide an explicit condition which, if confirmed empirically, proves that treatment-effect is not uniform, but varies appreciably across individuals.

Keywords: Heterogeneity, treatment on the treated, negative selection, effect modification, variable-effect bias

1 Introduction

Many social and health researchers are concerned with "the problem of heterogeneity," namely, the presence of idiosyncratic groups that react differently to treatment or policies. (Angrist, 1998; Angrist and Krueger, 1999; Elwert and Winship, 2010; Heckman and Robb, 1985; Heckman et al., 2006; Morgan and Todd, 2008; Morgan and Winship, 2007, 2015; Winship and Morgan, 1999; Xie et al., 2012). The reason is obvious. Health scientists need to know whether an approved drug is uniformly beneficial or kills some and saves more. Social scientists need to know whether those who have access to a program benefit most from the program; the alternative calls for revising recruiting policies (Brand and Xie, 2010).

Heterogeneity also introduces bias if one ventures to estimate average effects using linear or constant-effect models. Indeed, the bulk of the literature on this topic is concerned with demonstrating or minimizing this bias. Such bias is of no concern,

however, to students of nonparametric models where heterogeneity is assumed a priori within the model, thus protecting analysts from ever drawing conclusions that heterogeneity could invalidate.

Instead, nonparametric analysis concerns the detection of heterogeneity, if such exists, and locating its boundaries as narrowly as possible, within the granularity of the model. A straightforward way of assessing heterogeneity is to estimate the "interaction" or "effect modifying" capacity of various features of units. (VanderWeele and Robins, 2007). This amounts to estimating and comparing c-specific, or "conditional" effects, where C stands for a set of baseline covariates that characterize the units (Shpitser and Pearl, 2006).

This paper shows, however, that, under certain conditions, it is possible to assess the degree of heterogeneity in the population even without knowing the covariates C that make units differ in their response to treatment. We call this type of exogeneity "latent."

Section 2 of this paper will describe covariate-specific methods of detecting heterogeneity, and will summarize the capabilities and limitations of these methods. Section 3 defines a latent heterogeneity that produces differences between treated and untreated units. Section 4 will identify three settings in which this type of heterogeneity can be detected and assessed from empirical data. These include:

1. Randomized trials with binary treatments (Section 4.1)
2. Covariate adjustment (Section 4.2), and
3. Mediating instrumental variables (Section 4.3).

Section 5 presents a numerical example involving enrollment disparity in a job training program, where individuals possessing an unusual talent (a latent characteristics) have higher propensity to enroll in the program and are less likely to benefit from it. The section shows how the tests developed in Secions 4.1 and 4.2 can be used to detect such unusual characteristic and to assess its prevalent in the population.

Finally, Appendix A demonstrates the detection of a more drastic type of heterogeneity, where the population is composed of two distinct subpopulations, undetected by any observed characteristics, only through their behavior under both obsrvational and experimental studies. Appendix B will illustrate how structural models facilitate the evaluation of counterfactuals in genearl and heterogeneity in particular.

2 Covariate-induced Heterogeneity

If we can measure any characteristic C of individuals, a straightforward way of searching for heterogeneity is to determine if people having this characteristic respond differently from those not having it. There can of course be many group differences that escape measurement, this is unavoidable, but finding an observed characteristic accompanied by unusual effect size gives us a definitive warning that heterogeneity exists, and that its magnitude is at least equal to that found by examining C.

Formally, we can cast these considerations as follows.

2.1 Assessing covariate-induced heterogeneity

Let C stand for any measured baseline covariate, and let $E(Y_1 - Y_0|C = c)$ stand for the causal effect[1] in stratum $C = c$ of C. If $E(Y_1 - Y_0|C = c)$ is identifiable (for all c), we can then estimate the effect difference,

$$D(c_i, c_j) = |E(Y_1 - Y_0|C = c_i) - E(Y_1 - Y_0|C = c_j)| \tag{1}$$

for any two strata c_i and c_j of C. $D(c_i, c_j)$ gives the extent to which the effect size in group $C = c_i$ differs from that of group $C = c_j$. Further generalizing to all pairs (c_i, c_j), we get a lower bound LB on the heterogeneity between any two labeled groups in the population:

$$LB = max_{(c_i,c_j)} D(c_i, c_j) \tag{2}$$

This bound extends, of course, to the case where C is a vector of measured covariates and c_i, c_j any two instantiations of the variables in that vector. If we remove the requirement of identifiability, LB represents the best measure of heterogeneity in the population given the crudeness of our measurements. When the identifiability requirement is imposed, LB represents the best assessment of heterogeneity given both the crudeness of measurements and the opacity of non-experimental data. The two main problems in computing the lower bound in (2) is, first, to find a C for which the c-specific effect is identifiable and, second, to perform the maximization in (2) over all pairs (i, j) and all vectors C.

2.2 Special cases

Three special cases of estimable covariate-based heterogeneity are worth mentioning.

C is admissible[2]

If C is admissible, the c-specific effect is identified through

$$E(Y_1 - Y_0|C = c) = E(Y|X = 1, C = c) - E(Y|X = 0, C = c)$$

and $D(c_i, c_j)$ is estimable by simple regression.

[1] In this section we assume a binary treatment variable $X = (0, 1)$ and an outcome variable Y with two potential outcomes, Y_0 and Y_1, designating the hypothetical values of Y under treatment conditions $X = 0$ and $X = 1$, respectively.

[2] By "admissible" we mean a set C of covariates that satisfies the back-door criterion (Pearl, 1993; Pearl, 2009, pp. 79–81) in the causal diagram and thus permits the identification of the average causal effect by controlling for C. Admissibility entails the conditional independence $(Y_x \perp\!\!\!\perp X|C)$, sometimes called "conditional ignorability" (Rosenbaum and Rubin, 1983). The back-door criterion provides a scientific basis and a transparent test for "conditional ignorability" type claims, which many researchers entrust to intuition.

C is part of an admissible set

Assume C in itself is not admissible, but we can observe a set S of covariates such that $S \cup C$ is admissible (as in Fig. 1(b) and (c)). In such a case, the c-specific effect is still identifiable with:

$$E(Y_1 - Y_0|C = c) = \sum_s [E(Y|X = 1, S = s, C = c) - E(Y|X = 0, S = s, C = c)]P(s|c)^3$$

Figure 1 depicts four models in which the c-specific effect is identifiable, and two models in which it is not identifiable.

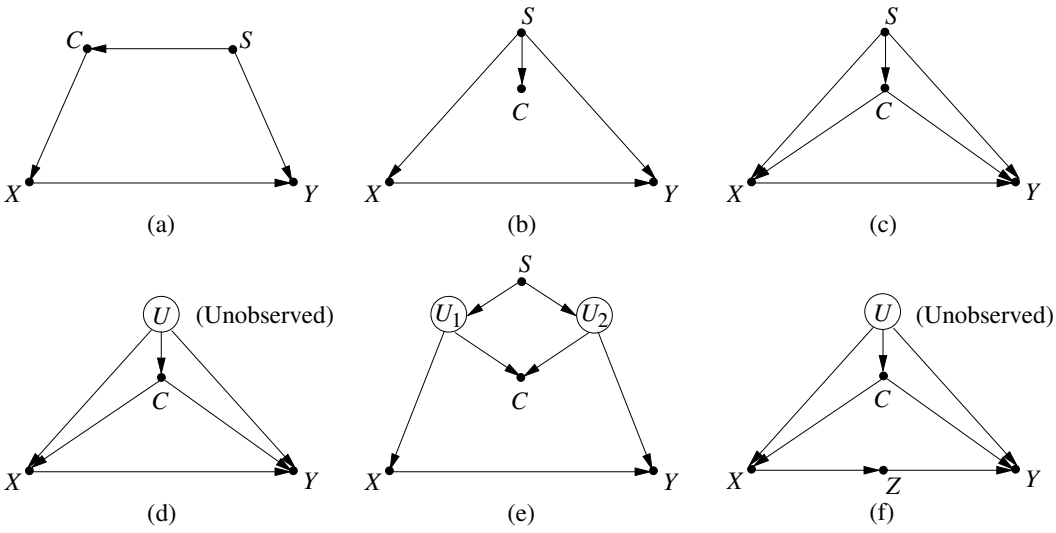

Figure 1: Models (a), (b), and (c) permit the identification of the c-specific effect of X on Y (by adjustment). Model (d) does not permit this identification, lacking an admissible set. Model (e) does not permit the identification of c-specific effects, even though S is admissible. Model (f) permits the identification using measurement of Z though S is admissible (U_1 and U_2 are unobserved).

Identification in the absence of admissible sets

If C is not part of an admissible set, the c-specific effect cannot be identified by adjustment. A typical example is given in Fig. 1(d). Since U is unobserved, the confounding path $X \leftarrow U \rightarrow Y$ remains open even if we adjust for C. However, the measurement of other variables in the model may nevertheless permit the identification of $E(Y_1 - Y_0|C = c)$ by other methods, and the bound LB can be estimated accordingly. An example is given in Fig. 1(f) where $E(Y_1 - Y_0|C = c)$ is identifiable through the front-door estimator (Pearl, 1995, see also Section 4.3) by virtue of measuring an intermediate variable Z. A complete characterization of models that permit the identification of c-specific effects is given by Shpitser and Pearl (2006).

[3]In practice, the summation over S can be prohibitive, and propensity score methods can be used to replace this summation by integration over the unit interval $0 \leq PS \leq 1$ (**?**)

C excluded from all admissible sets

An intriguing pattern of heterogeneity is described in Fig. 1(e). Here S is an admissible set, but if we add C to S, admissibility is destroyed. This occurs because C is a collider, so conditioning on C would open the path $X \leftarrow U_1 \rightarrow C \leftarrow U_2 \rightarrow Y$ in violation of the back-door condition. This means that, even if C is observed, we cannot identify the c-specific effects (of X on Y) and, therefore, we cannot assess whether units falling in different strata of C differ in their response to X. Adjustment for c_i or c_j, be it with or without S, would tell us nothing about the causal effects in those strata, and would thus prevent us from using the comparisons described in Section 2.1, Eq. (1).

Note that Model (e) is statistically indistinguishable from (c), implying that no statistical test, however clever, can determine whether a given set $\{S, C\}$ of covariates is admissible. This includes sensitivity analysis which is often presumed to provide evidence for ignorability or admissibility.

3 Latent Heterogeneity Between the Treated and Untreated

So far, the aim of the analysis has been to find two subgroups $C = c_i$ and $C = c_j$ with unequal effect sizes, where C was an observed baseline characteristic of individuals. In this section we abandon this requirement and seek "latent heterogeneity," namely, heterogeneity that is not present in any baseline covariate but stems from unknown origin and manifests itself in effect differences between the treated and untreated groups.

3.1 Two types of confounding

The potential for detecting such heterogeneity was unveiled in the analyses of Winship and Morgan (1999) and Xie et al. (2012), who decomposed the *average treatment effect* ATE into several components:[4]

$$ATE = E(Y_1 - Y_0) = E(Y|X=1) - E(Y|X=0) - [E(Y_0|X=1) - E(Y_0|X=0)] \\ - (ETT - ETU)/P(X=0)$$

[4]This decomposition was first proposed in sociology by Winship and Morgan (1999, p. 667) in a paper that raised awareness for the importance of treatment-effect heterogeneity. Emphasis on ETT and ETU was introduced earlier in econometrics by Heckman and his co-workers (Heckman and Robb, 1986; Heckman, 1992).

where ETT and ETU are the average *effect of treatment on the treated* and untreated respectively[5] i.e.,

$$ETT = E(Y_1 - Y_0|X = 1),$$
$$ETU = E(Y_1 - Y_0|X = 0).$$

They observed that the bias,

$$Bias = E(Y|X = 1) - E(Y|X = 0) - ATE$$

is made up of two components with distinct characteristics. The first is $[E(Y_0|X = 1) - E(Y_0|X = 0)]$ and the second is $ETT - ETU$. The former is not a causal effect but merely a difference in output (Y) between two groups under the same "no-treatment" regime. The latter, on the other hand, represents difference in treatment effects of two groups, the treated and the untreated, and would be non-zero only if the two groups respond differently to treatment, thus exhibiting heterogeneity.[6]

Xie et al. called the former Type-I bias and the latter Type-II bias, whereas Morgan and Winship (2007, pp. 46–8) called them *baseline bias* and *differential treatment effect bias*. We will shorten the labels to read *baseline* and *variable-effect* biases respectively. To understand the two types of bias, think about two groups, one with high Y that is aggressively selected for treatment, and one with low Y, which is rarely selected for treatment. There will definitely be a bias in estimating ATE, even if all units have the same treatment effect. Now think about two other groups, both achieving the same Y under no treatment, but one is sensitive to X and one is not. If the second is more likely to select treatment, a bias is generated solely by the sensitivity difference between the two groups.

3.2 Separating fixed-effect from variable-effect bias

To convince ourselves that baseline and variable-effect biases, as defined above, indeed capture fixed-effect and variable-effect subpopulations, respectively, we evaluate their corresponding expressions in a linear model with an interaction term. The model is shown in Fig. 2 and represents the structural equations:

$$y = \beta x + \gamma z + \delta xz + \epsilon_1$$
$$x = \alpha z + \epsilon_2$$
$$z = \epsilon_3$$

where the disturbances ϵ_1, ϵ_2, and ϵ_3 are assumed to be mutually independent. Indeed, for variable-effect bias we obtain:[7]

$$ETT - ETU = \alpha\delta(x' - x)^2$$

[5]Xie et al. (2012) used D for treatment and $TT - TUT$ instead of $ETT - ETU$. In contrast, Morgan and Winship (2015) use $ATT - ATC$. Here we use X for treatment, consistent with theoretical analyses in Shpitser and Pearl (2009), where the acronym ETT was used, and a necessary and sufficient condition for identifying ETT was developed.

[6]Heckman et al. (2006) called this difference *essential heterogeneity*.

[7]These expressions follow directly from the structural definition of counterfactuals (Pearl, 2009, p. 98) as defined in Eq. (12). A complete derivation is given in Appendix B.

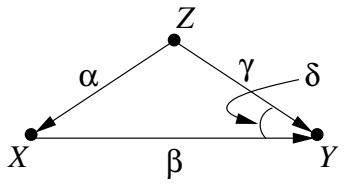

Figure 2: A linear model with interaction, demonstrating baseline and variable-effect biases. The former is proportional to $\gamma\alpha$ and independent of δ; the latter is proportional to $\delta\alpha$ and independent of γ, reflecting effect variability.

whereas for baseline bias we have

$$E(Y_x|X = x') - E(Y_x|X = x)] = \gamma\alpha(x' - x).$$

(x and x' are two arbitrary levels of the treatment.) This is exactly the decomposition we expect; the former captures the bias introduced through the interaction term δ (representing variable-effect), whereas the latter represents the bias that would prevail in the linear (or fixed-effect) case, without that interaction.

Note also the $ETT - ETU$ vanishes when $\alpha = 0$. Thus, not every effect heterogeneity is detected through the difference $ETT - ETU$. When interactions are strong (i.e., high δ) we certainly have appreciable heterogeneity between units with high Z and units with low Z. However, this heterogeneity will remain undetected; it will not be revealed through the difference $ETT - ETU$, unless Z also affects the treatment assignment X.

4 Three Ways of Detecting Heterogeneity

The interesting feature in the preceding analysis is that the decomposition into fixed-effect and variable-effect components can be defined counterfactually, without resorting to a specific model or a specific covariate set. This means that whenever we can identify ETT and ETU, we can also obtain an indication of heterogeneity, regardless of whether we can name or observe the covariates responsible for the heterogeneity. Moreover, even in cases where auxiliary measurements are needed for identifying ETT and ETU, the graphical theory of ETT (Shpitser and Pearl, 2009) can guide us in the assessment of heterogeneity by (1) selecting the right set of measurements and (2) obtaining the right estimands for ETT and ETU.

The three classical cases where ETT can be identified are:

1. The treatment is binary, and $E(Y_1)$ and $E(Y_0)$ are identifiable by some method (e.g., randomized trials).

2. The treatment is arbitrary, and $E(Y_x)$ is identifiable (for all x) by adjustment for an admissible set of covariates.

3. ATE is identified through mediating instruments.

The next subsections deal separately with each of these cases.

4.1 Detecting heterogeneity in randomized trials

It is well known that, when treatment is binary, ETT and ETU are identified whenever $E(Y_0)$ and $E(Y_1)$ are (Pearl, 2009, p. 396–7). Moreover, the relation between these quantities is given by

$$\begin{aligned} ETT &= E(Y_1 - Y_0|X = 1) = \\ &= E(Y|X = 1) - [E(Y_0) - E(Y|X = 0)(1 - p)]/p \\ ETU &= E(Y_1 - Y_0|X = 0) \\ &= [E(Y_1) - E(Y|X = 1)p]/(1 - p) - E(Y|X = 0) \end{aligned}$$

where $p = P(X = 1)$.[8]

We conclude that in a (binary) randomized clinical trial, where $E(Y_0)$ and $E(Y_1)$ are estimable empirically, the difference $ETT - ETU$ is estimable as well and is given by

$$ETT - ETU = [E(Y|X = 1) - E(Y_1)]/(1 - p) + [E(Y|X = 0) - E(Y_0)]/p \quad (3)$$

Likewise, the size of the baseline bias is identifiable from clinical trials, and is given by:

$$E(Y_0|X = 1) - E(Y_0|X = 0) = [E(Y_0) - E(Y|X = 0)]/p \quad (4)$$

This means that, based on pre-trial and post-trial data we can estimate the heterogeneity bias that exists in the population prior to randomization, and we can accomplish this without measuring any covariate whatsoever.

This result might appear surprising at first; how can we possibly detect the existence of individual variations among units when we have only population data? Upon further reflection, however, we note that $ETT - ETU$ does not represent the degree of heterogeneity in the population but rather that portion of heterogeneity that exhibits preferential selection to treatment. Additionally, we are not entirely justified in claiming that we accomplish this assessment without measuring *any* covariate. The treatment itself serves as a measured covariate in our case, since it is a proxy for those factors that affect the choice of treatment.

While these explanations mitigate the surprise, the point remains that effect heterogeneity is not entirely shielded from empirical scrutiny, even when we only have population data. Whenever experimental findings reveal a non-zero $ETT - ETU$, one can categorically state that heterogeneity exists in the population, that is, there exist at least two groups whose treatment effects differ from one another.

The analysis also tells us which combination of observational and experimental data would compel us to conclude that the population consists of at least two disparate groups. In particular, Eq. (3) implies that whenever we observe the inequality

$$P(X = 1)[E(Y|X = 1) - E(Y_1)] \neq P(X = 0)[E(Y|X = 0) - E(Y_0)] \quad (5)$$

[8]These expressions can readily be derived by noting that $E(Y_0|X = 0) = E(Y|X = 0)$ and writing:

$$E(Y_0) = E(Y_0|X = 1)p + E(Y|X = 0)(1 - p).$$

For non-binary treatments, ETT is not expressible in terms of $E(Y_0)$ and $E(Y_1)$.

we can be assured that the population is marred by heterogeneity, and in such cases, a systematic exploration may be undertaken to unveil its underlying sources. This is not a trivial result by any means; it is in fact counter intuitive, and should be considered a victory of formal counterfactual analysis. Section 5 presents numerical examples of such findings, and Appendix A provides an example where Eq. (5) returns equality despite rampant heterogeneity.

Sander Greenland suggested (personal communication) that heterogeneity in randomized trials is related to the issue debated by Fisher vs. Neyman about the appropriate nulls to test. Fisher advocated the strict (point) null $Y_1 = Y_0$ for all units, (which led to his famous exact test); in contrast, Neyman advocated the much weaker mean null $E(Y_1) = E(Y_0)$, which allows arbitrarily extensive heterogeneity, ostensibly on the grounds that nothing finer could be discerned in a randomized experiment (Greenland, 1991).

Equation (5) casts this debate in a new setting. While Fisher's exact null cannot be distinguished from Neyman's mean null in a pure randomized experiment, such distinction is feasible when we have a combination of randomized and observational data. In fact, inequality in Eq. (5) can be regarded as a testable condition for rejecting Fisher's null hypothesis.

Section 5 and Appendix A present models where Neyman's mean null holds, $E(Y_1) = E(Y_0)$, as well as inequality in (5), thus rejecting Fisher's sharp null. The same test can be applied when the outcome distribution under treatment is identical to the outcome distribution for control, a case where conventional approaches to testing heterogeneity fail (Greenland, 1999; Ding, 2014).

4.2 Detecting heterogeneity through adjustment

The second case where *ETT* and *ETU* are identified is when an admissible set Z of covariates can be measured, yielding (see footnote 2) the adjustment estimand

$$E(Y_x) = \sum_z E(Y|x,z)P(z) \tag{6}$$

where x is any treatment level, not necessarily one or zero. It can be further shown that if Z is admissible, the expression for $E(Y_x|x')$ can be identified as well, and is given by

$$E(Y_x|x') = \sum_z E(Y|x,z)P(z|x') \tag{7}$$

(Shpitser and Pearl, 2009). It is almost the same as the adjustment formula (6), save for using $P(z|x')$ as a weighting function, instead of $P(z)$.[9]

[9]This difference accounts for the modified Horvits-Thompson weights required for estimating *ETT* and *ETU* by regression (Morgan and Winship, 2015, p. 231).

Accordingly, we can write the difference $ETT - ETU$ as

$$\begin{aligned} ETT - ETU &= E(Y_{x'} - Y_x|X = x') - E(Y_{x'} - Y_x|X = x) \\ &= \sum_z [E(Y|X = x', z) - E(Y|X = x, z)][P(z|X = x') - P(z|X = x)] \end{aligned} \tag{8}$$

and thus establish an explicit and general formula for the detectable part of variable-effect heterogeneity.[10]

When the set Z is large, the estimation of (8) can be enhanced using propensity score adjustment. But aside from providing a powerful estimation method in sparse data studies, the use of propensity scores does not add any substance to the discussion of identification (Pearl, 2009, pp. 348–52).

An objection might be raised to classifying the heterogeneity detected by Eq. (8) as "latent" when, in fact, it could only be uncovered using a set Z of observed covariates. The justification rests on the realization that the treated-untreated heterogeneity, $ETT - ETU$, is a property of the population, not of the set Z chosen to uncover it. Z serves merely as an auxiliary tool for uncovering $ETT - ETU$; it does not affect its value. Moreover, $ETT - ETU$ represents a new species of heterogeneity, unrelated to those induced by the strata of Z (see Section 2.2). To witness, Eq. (8) shows that the heterogeneity between the treated and untreated groups may be many times larger than that induced by any two strata of Z. For a trivial, albeit contrived example, let Z take on integer values $z = 1, 2, \ldots, k$, and let

$$E(Y|X = x', z) - E(Y|X = x, z)$$

be positive for even values of z and negative for odd values. If we now let the difference $P(z|X = x') - P(z|X = x)$ be positive for even values and negative for odd values of z, $ETT - ETU$ increases indefinitely as k increases, while the effect difference between any two strata of Z remains bounded. We also note, somewhat counterintuitively, that the treated-untreated heterogeneity ($ETT - ETU$) vanishes within each stratum $Z = z$ of an admissible set Z, while the overall difference $ETT - ETU$ need not be zero. The reason is that ETT and ETU invoke different weighing functions in averaging over the values of z; $P(z|X = x')$ is invoked in the former and $P(z|X = x)$ in the latter.[11]

[10] Morgan and Todd (2008) recognized the fact that ETT and ETU are estimable (using weighted regression) whenever conditional ignorability holds. Equation (8) extends their analysis by providing an explicit formula for $ETT - ETU$, applicable whenever a set Z of covariates is observed that is deemed admissible for identifying ATE. (Note that identifying ATE, in itself, is insufficient.) Brand and Halaby (2005) used bootstrapping methods to determine whether the difference between the ETT and the ETU is significant?

[11] This is an interesting variant of Simpson's paradox that surfaces when the aggregation of data results in sign reversal of all statistical associations (Blyth, 1972; Simpson, 1951). However, in the standard exposition of Simpson's paradox, the signs of all causal effects remain unaltered (Pearl, 2009, pp. 180–2; 2014). Here we witness a causal, not associational relationship that is present in the combined population and is absent in each and every subpopulation.

4.3 Detecting heterogeneity through mediating instruments

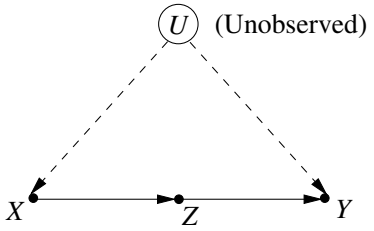

Figure 3: A model in which variable Z acts as a mediating instrument for identifying the causal effect of X on Y in the presence of unknown or unobserved confounders (U).

Identification by adjustment requires modeling assumptions that researchers may not be prepared to make. Attempting to circumvent this requirement, some researchers have advocated the use of instrumental variables (IV), which appears to require milder assumptions (Angrist and Pischke, 2010). Aside from the fact that good instruments are hard to come by, and that the choice of instruments often requires strong modeling assumptions, identification through instruments suffers from a fundamental limitation in that it is effective only in linear (or pseudo-linear) models, and in nonparametric models, can only identify local effects, sometimes called $LATE$ (Angrist et al., 1996; Brand and Thomas, 2013).

Fortunately, the use of mediating instruments overcomes these limitations and identifies causal effects in non-parametric models even in the presence of unknown confounders. The method of mediating instruments, also known as "the front-door criterion" (Pearl, 1995) is depicted in Fig. 3, and assumes the availability of covariates Z that intercept all directed paths from treatment (X) to outcome (Y).[12] Moreover, the graphical theory of ETT teaches us that both ETT and ETU are identifiable in the model of Fig. 3 and can be obtained from the estimand:

$$E(Y_x|X = x') = \sum_z E(Y|z, x')P(z|x) \tag{9}$$

where x and x' are any two levels of the treatment (Shpitser and Pearl, 2009).

Remarkably, this expression is almost identical to the one obtained through adjustment for confounders Z, Eq. (7), save for exchanging x and x'. Moreover, and in contrast to identification by randomized experiment, this estimand remains valid for non-binary treatments as well.

Accordingly, the estimand for the heterogeneous component of the bias becomes

[12]For application of the front-door criterion in the social sciences, see (Chalak and White, 2012; Morgan and Winship, 2007, 2015).

identical to that of Eq. (8):

$$\begin{aligned} ETT - ETU &= E(Y_{x'} - Y_x|X = x') - E(Y_{x'} - Y_x|X = x) \\ &= \sum_z [E(Y|X = x', z) - E(Y|X = x, z)][P(z|X = x') - P(z|X = x)] \end{aligned} \tag{10}$$

with $X = x'$ representing the treatment level received and $X = x$ a comparison reference. Likewise, the expression for the baseline component of the bias becomes:

$$E(Y_x|X = x') - E(Y_x|X = x) = \sum_z [E(Y|z, x') - E(Y|z, x)]P(z|x) \tag{11}$$

We are now in possession of simple expressions for both the heterogeneous and homogeneous parts of the bias. These expressions enable us to decompose the bias into its heterogeneous and homogeneous parts without any reference to the latent confounders (U), which may remain unknown or unnamed. Whereas detection by randomized trials requires physical control, and is limited to binary treatments, and detection through ordinary adjustment requires an admissible set of deconfounders, the method of mediating instruments gives us a general way of assessing the impact of homogeneous vs. heterogeneous mechanisms on the observed bias without knowing the actual mechanisms involved.

5 Example: Heterogeneity in Recruitment

A government is funding a job training program aimed at getting jobless people back into the workforce. A pilot randomized experiment shows that the program is effective; a higher percentage of people were hired among those trained than among the untrained. As a result, the program is approved, and a recruitment effort is launched to encourage enrollment among the unemployed.

A study conducted a year later reveals that the hiring rate among the trained is even higher than in the randomized study. Still, critics claim that the program is a waste of tax payers' money because, while the program was somewhat successful in the experimental study, where participants were chosen at random, there is no proof that the program accomplishes its mission among those recruited for enrollment. Those enrolled, so the critics say, are more intelligent, more resourceful, and more socially connected than the eligibles who did not enroll, and would have found a job regardless of the training. The population is not homogeneous, the critics claim; the informed who are first to enroll draw little benefit from the program, while the weak and uninformed who could truly benefit from it were not aggressively recruited.

In order to assess the extent to which the $ETT - ETU$ test can detect the presence of such heterogeneity, we will simulate the hiring process assuming two types of individuals, "informed" and "uninformed." Let $Z = 1$ stand for the class of "informed" individuals, for whom the chances of hiring after training is only 10% higher than without training, 0.9 vs. 0.8. Let Z=0 stands for the class of uninformed individuals, for whom the chances of hiring after training are 70% higher

than without training, 0.8 vs. 0.1. We will assume that the propensity for enrollment among the informed, q_2, is higher then that among the uninformed, q_1, i.e., $q_2 - q_1 = P(X = 1|Z = 1) = P(X = 1|Z = 0) > 0$.

Since we are dealing with a binary treatment, we can assess the magnitude of $ETT - ETU$ using Eq. (3) without measuring any covariates. We rely solely on $\{E(Y_1), E(Y_0)\}$, which are estimable from the experimental study and $\{E(Y|X = 1), E(Y|X = 1)\}$, which are estimable from the observational study, and reflect the current recruitment policy. The plots in Fig. 4 depict the difference $ETT - ETU$ as a function of r, the percentage of "informed" individuals in the population, with each curve representing a fixed enrollment disparity $q_2 - q_1$.

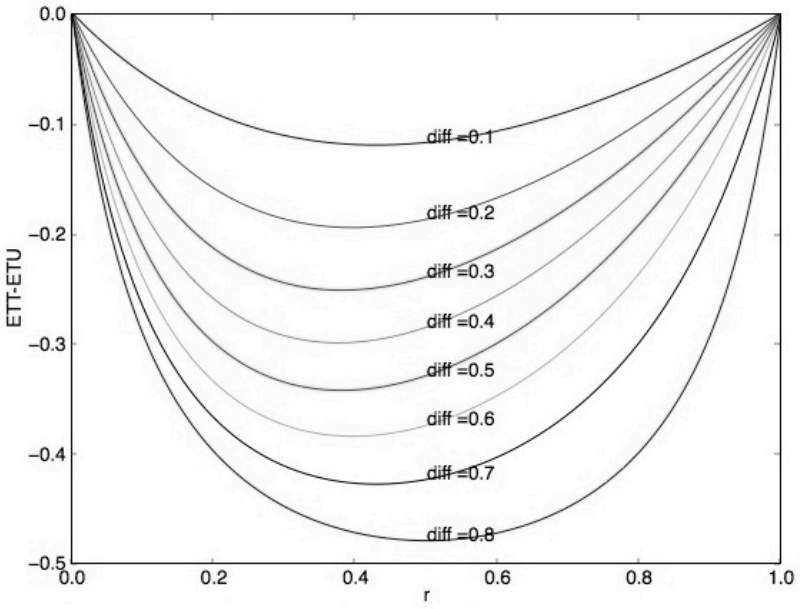

Figure 4: $ETT - ETU$ vs. r for different levels of enrollment disparity, $q_2 - q_1$.

In generating these plots, we assume a model similar in structure to the one of Fig. 2, with Z being the only confounder between X and Y. We further assume the

following parameters:

$$\begin{aligned} E[Y|X=1,Z=1) &= 0.9 \\ E[Y|X=0,Z=1) &= 0.8 \\ E[Y|X=1,Z=0) &= 0.8 \\ E[Y|X=0,Z=0) &= 0.1 \\ q_1 = P(X=1|Z=0) &= 0.1 \end{aligned}$$

We see that $ETT - ETU$ is negative, indicating loss of opportunity due to misdirected recruiting policy, with those in the program benefitting less from it than (potentially) those who are not in it. The higher the enrollment discrepancy $q_2 - q_1$ between the "informed" and the "uninformed," the more negative the difference $ETT - ETU$. We further see that the difference $ETT - ETU$ becomes zero when the population becomes homogeneous, at $r = 0$ or $r = 1$, with the slopes at these two points measuring the sensitivity of program effectiveness to the presence of heterogenous individuals. Plots such as those in Fig. 2 provide valuable information about the nature and magnitude of the heterogeneity observed. For example if in a randomized experiment we observe the difference $ETT - ETU = -0.3$ (through Eq. (3)), we can then infer that, if the propensity difference $q_1 - q_2$ is lower than 0.5, the proportion r must lie between 0.20 and 0.62. The larger the difference $q_1 - q_2$ the wider the bounds for r.

6 Conclusions

This paper explores ways of uncovering the presence of effect-heterogeneity without knowing the factors that may produce it. This possibility was shown to be realizable in the three most common designs in which the average treatment effect (ATE) can be estimated: (1) randomized experiments, (2) covariate adjustment, and (3) mediating instruments. The only exceptions in these three designs are randomized experiments with non-binary treatments, and models in which ATE is identified and ETT is not. Such models can be recognized using the graphical theory of ETT (Shpitser and Pearl, 2009), which provides a complete set of conditions for the identification of ETT and ETU from modeling assumptions.

In all three cases that allow for the detection of latent heterogeneity, we have derived explicit conditions that, if observed in practice, behoove us to conclude that subpopulations exist that differ in their response to treatment. These conditions can also serve to assess, albeit roughly (in the form of lower bounds), the magnitude of the heterogeneity detected.

Acknowledgment

I am indebted to Jennie Brand and Steven Morgan for calling my attention to the sociological literature on heterogeneity and commenting on earlier versions of the manuscript. Subsequently, the paper benefitted from discussions with Felix Elwert and Sander Greenland. I thank Ang Li for generating the plots of Fig. 4.

Appendix A (An Extreme Case of Latent Heterogeneity)[13]

The example below demonstrates a case in which the bias is zero, the average causal effect is zero and, yet, heterogeneity is high and can be detected by Eq. (5), using no modeling assumptions.

A study was conducted to determine which of two schools, A or B, has a more effective educational program. 200 randomly selected students underwent a randomized trial and were randomly assigned to the two schools, 100 to each. Another group of 200 (randomly selected) students were allowed to choose schools on their own; 100 selected A and 100 B. After a year of study, students were tested in a uniform, state run exam, and data showed the following:

100% of the A-choosing students failed the state exam

100% of the B-choosing students failed the state exam

50% of the A-randomized students failed the state exam

50% of the B-randomized students failed the state exam

It appears that, when given a choice, students tend to pick the school that is worse for them, which is strange but explainable. Suppose school A deemphasized math and B deemphasized history, while the state exam demands proficiency in both math and history. If students choose schools by the area of their strength, then free choice amounts to a license to neglect one of the required subjects, which is a ticket to failure. Random assignment would force at least 50% of the students to study an area of weakness, which may explain the 50% success rate in the randomized groups.

From the data available, and letting $X = 1$ and $X = 0$ stand for "School A chosen" and "School B chosen," respectively, we can infer the following findings:

$$\begin{array}{lll} p = \frac{1}{2}, & E(Y|X=1) = 0, & E(Y|X=0) = 0 \\ & E(Y_1) = \frac{1}{2}, & E(Y_0) = \frac{1}{2} \end{array}$$

[13]This example is taken from (Pearl, 2012).

Accordingly we have:

$$\begin{aligned} \textit{Bias} &= E[(Y|X=1) - E(Y|X=0)] - [E(Y_1) - E(Y_0)] \\ &= 0 - 0 - (\frac{1}{2} - \frac{1}{2}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \textit{Baseline Bias} &= E(Y_0|X=1) - E(Y_0|X=0) \\ &= [E(Y_0) - E(Y|X=0)]/p \\ &= (\frac{1}{2} - 0)2 \\ &= 1 \end{aligned}$$

$$\begin{aligned} \textit{Variable-effect Bias} &= (ETT - ETU)(1-p) \\ &= [E(Y|X=1) - E(Y_1)]p/(1-p) + [E(Y|X=0) - E(Y_0)]] \\ &= (0 - \frac{1}{2}) + (0 - \frac{1}{2}) \\ &= -1 \end{aligned}$$

We conclude that a substantial effect-heterogeneity exists in the population. In fact, the bias is composed of two components of equal magnitude and opposite sign. This result is not surprising given that our population is composed indeed of two distinct subpopulations, indexed by school preference, which have two different treatment effects. Those who prefer school B have clearly different benefit from A vs. B as compared to those who prefer school A; the former would pass the exam, the latter would fail.

It is also interesting, at this point, to examine models in which latent heterogeneity is rampant, yet remains undetected by the difference $ETT - ETU$. Such models are discussed in (Pearl, 2009, pp. 35–6), which can be adapted to the story above by assuming that Z (students school preference) is totally independent of X (the school actually attended). In such an environment, the two groups will still exhibit the disparate treatment effects, but the difference $ETT - ETU$ will be zero, because the relationship between X and Y is not confounded.

Appendix B (Assessing Heterogeneity in Structural Equation Models)

In this Appendix, I first define counterfactuals in terms of structural equation models, and then illustrate how this definition facilitates the detection of heterogeneity in the linear model discussed in Section 3.2. The definition is fundamental to the understanding of counterfactuals in general, and for that reason, I will first introduce the method and then solve the example in minute details. The solution will demonstrate the role of structural models in defining and evaluating counterfactuals.

B.1 The Structural origin of counterfactuals

At the center of the definition lies a model M consisting of a set of equations that represents the investigator's perception of reality. M consists of two sets of variables, U and V (exogenous and endogenous), and a set F of equations that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u, of all variables in V and U and accordingly *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables corresponds to defining a "unit," or a "situation" in the model, and uniquely determines the values of all variables in V. Therefore, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V. The probabilities on U and V can best be interpreted as the proportion of the population with a particular combination of values on U and/or V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$, where Y and X are any variables in V. The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x" as an instruction to make a minimal modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X with a constant x, which may be thought of as an external intervention $do(X = x)$, not necessarily by a human experimenter, that imposes the condition $X = x$. This replacement permits the constant x to differ from the actual value of X (namely $f_x(v, u)$) without rendering the system of equations inconsistent, thus allowing all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M, with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{12}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x.

This definition, first proposed in (Balke and Pearl, 1994a,b) was recently dubbed the "First Law of causal inference" (Pearl, 2015) due to its universality, and because it treats counterfactuals as an intrinsic property of reality rather than a byproduct of a specific experimental design. Simon and Rescher (1966) came close to this definition but, lacking the "wiping out" operator, could not reconcile the contradiction that evolves when an observation $X = x'$ clashes with the antecedent $X = x$ of the counterfactual Y_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models. (see also Pearl, 2009, Chapter 7). They showed that the axioms governing recursive structural counterfactuals are identical to those used in the

potential outcomes framework, hence the two systems are logically identical – a theorem in one is a theorem in the other. This means that relying on structural models as a basis for counterfactuals does not impose additional assumptions beyond those routinely invoked by potential outcome practitioners.

$P(u)$ induces a well defined probability distribution on V, $P(v)$. As such, it not only defines the probability of any single counterfactual, also assigns joint distribution of all conceivable counterfactuals, including those that may not be observed. Thus the probability of the Boolean combination, "$Y_x = y$ AND $Z_{x'} = z$" for variables Y and Z in V and two different values of X, x and x', is well-defined even though it is impossible for both outcomes to be simultaneously observed as $X = x$ and $X = x'$ cannot be concurrently true.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any information about an individual, can be computed by the 3-step process:

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

B.2 Illustration

To demonstrate the power of this definition, let us compute the latent heterogeneity $ETT - ETU$ for the interaction model discussed in Section 3.2. The model (shown in Fig. 2) represents the structural equation model:

$$\begin{aligned} M: \quad & Y = \beta X + \gamma Z + \delta XZ + \epsilon_1 \\ & X = \alpha Z + \epsilon_2 \\ & Z = \epsilon_3 \end{aligned}$$

The modified model M_x, representing the intervention $X = x$, is given by

$$\begin{aligned} M_x: \quad & Y = \beta X + \gamma Z + \delta xZ + \epsilon_1 \\ & X = x \\ & Z = \epsilon_3 \end{aligned}$$

Let $X = x$ represent the treatment administered and $X = x'$ the level that X attains under natural, no-treatment conditions. We first compute the conditional counterfactual $E(Y_x|X = x')$ which appears in the expressions of ETT and ETU

$$\begin{aligned} ETT &= E[Y_x - Y_{x'}|X = x] \\ ETU &= E[Y_x - Y_{x'}|X = x'] \end{aligned}$$

Since Y_x is equal to the solution for Y in the mutilated model M_x, we have

$$\begin{aligned} E[Y_x|X = x'] &= E[\beta x + \gamma Z + \delta xZ + \epsilon_1|X = x'] \\ &= \beta x + (\gamma + \delta x)E[Z|X = x'] \end{aligned}$$

where we make use of the orthogonality assumption $\epsilon_1 \perp\!\!\!\perp X$. Further assuming standardized variables (i.e., zero mean and unit variance) we have $E[Z|X = x'] = \alpha x'$, which leads to

$$E[Y_x|X = x'] = \beta x + (\gamma + \delta x)\alpha x'$$

Accordingly, the effect of treatment on the treated is given by

$$\begin{aligned} ETT &= E[Y_x - Y_{x'}|X = x] \\ &= E[Y|X = x] - E[Y_{x'}|X = x] \\ &= \beta x + \alpha\gamma x + \alpha\delta x^2 - [\beta x' + (\gamma + \delta x')\alpha x] \\ &= (\beta + \alpha\delta x)(x - x') \end{aligned}$$

In a similar fashion we obtain

$$\begin{aligned} ETU &= E[Y_x - Y_{x'}|X = x'] \\ &= (\beta + \alpha\delta x')(x - x') \end{aligned}$$

and finally:

$$ETT - ETU = \alpha\delta(x - x')^2.$$

which confirms the result stated in Section 3.2.

References

Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association* **91** 444–472.

Angrist, J. D. (1998). Estimating the labor market on voluntary military service using social security date on military applicants. *Econometrica* **66** 249–288.

Angrist, J. D. and Krueger, A. B. (1999). Handbook of labor economics. In *Causality: Statistical Perspectives and Applications* (O. Ashenfelter and D. Card, eds.), edition 1, volume 3, chapter 23 ed. Elsevier, 1277–1366.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* **24** 3–30.

Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

Blyth, C. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* **67** 364–366.

Brand, J. E. and Halaby, C. N. (2005). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research* **35** 749–770.

Brand, J. E. and Thomas, J. S. (2013). Causal effect heterogeneity. In *Handbook of Causal Analysis for Social Research* (S. L. Morgan, ed.), chap. 11. Springer, Netherlands, 189–213.

Brand, J. E. and Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review* **75** 273–302.

Chalak, K. and White, H. (2012). An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics* **44** 1–31.

Ding, P. (2014). A paradox from randomization-based causal inference. Tech. rep., Harvard University, Cambridge, MA. arXiv:1402.0142v3.

Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, UK, 327–336.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *The American Statistician* **45** 248–251.

Greenland, S. (1999). Relation of probability of causation, relative risk, and doubling dose: A methodologic error that has become a social problem. *American Journal of Public Health* **89** 1166–1169.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Heckman, J. (1992). Randomization and social policy evaluation. In *Evaluations: Welfare and Training Programs* (C. Manski and I. Garfinkle, eds.). Harvard University Press, Cambridge, MA, 201–230.

Heckman, J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data* (J. Heckman and B. Singer, eds.). Cambridge University Press, New York, NY, 156–245.

Heckman, J. and Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inference From Self Selected Samples* (H. Wainer, ed.). Springer-Verlag, New York, NY, 63–107.

Heckman, J., Urzua, S. and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics* **88** 389–432.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press, New York, NY. 2nd edition, 2015.

Morgan, S. and Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. 2nd ed. Cambridge University Press, New York, NY.

Morgan, S. L. and Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* **38** 231–281.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

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

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science* **13** 426–436, DOI: 10.1007/s11121–011–0270–1.

Pearl, J. (2015). Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory* **31** 152–179. Special issue on Haavelmo Centennial.

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

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*. AUAI Press, Corvallis, Oregon, 514–521.

Simon, H. and Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science* **33** 323–340.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13** 238–241.

VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18** 561–568.

Winship, C. and Morgan, S. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology* **25** 659–706.

Xie, Y., Brand, J. E. and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology* **42** 314–347.