Article

Sociological Methods & Research
1-20

DOI: 10.1177/0049124115600597
smr.sagepub.com

Detecting Latent Heterogeneity

Judea Pearl[1]

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

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, Urzua, and Vytlacil 2006; Morgan and Winship 2007, 2015; Morgan and Todd

[1] Computer Science Department, University of California, Los Angeles, CA, USA

Corresponding Author:
Judea Pearl, Computer Science Department, University of California, Los Angeles, CA 90095, USA.
Email: judea@cs.ucla.edu

2008; Winship and Morgan 1999; Xie, Brand, and Jann 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 article 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."

The second section of this article will describe covariate-specific methods of detecting heterogeneity and will summarize the capabilities and limitations of these methods. The third section defines a latent heterogeneity that produces differences between treated and untreated units. The fourth section 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 (Detecting Heterogeneity in Randomized Trials subsection),
2. covariate adjustment (Detecting Heterogeneity Through Adjustment subsection), and
3. mediating instrumental variables (Detecting Heterogeneity Through Mediating Instruments subsection).

The fifth section 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 Detecting Heterogeneity in Randomized Trials and Detecting Heterogeneity Through Adjustment subsections can be used to detect such unusual characteristic and to assess its prevalence in the population.

Finally, Online 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 observational and experimental studies (Pearl 2013).[1] Online Appendix B will illustrate how structural models facilitate the evaluation of counterfactuals in general and heterogeneity in particular.

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.

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[2] 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:

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

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:

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

This bound extends, of course, to the case where C is a vector of measured covariates and c_i, c_j are 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 nonexperimental data. The two main problems in computing the lower bound in equation (2) are, first, to find a C for which the c-specific effect is identifiable and, second, to perform the maximization in equation (2) over all pairs (i, j) and all vectors C.

Special Cases

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

C is admissible. If C is admissible,[3] 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.

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 Figure 1b and c). In such a case, the c-specific effect is still identifiable with[4]:

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

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

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 Figure 1d. 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 Figure 1f, where $E(Y_1 - Y_0|C = c)$ is identifiable through the front-door estimator (Pearl 1995, see also Detecting Heterogeneity Through Mediating Instruments subsection) 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).

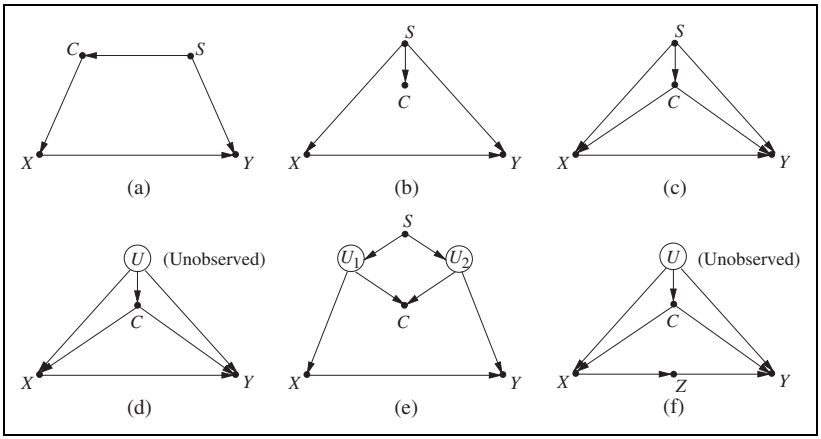

Figure 1. Models (a), (b), and (c) permit the identification of the *c*-specific effect of *X* on *Y* (by adjustment). Model (d) does not permit this identification, lacking an admissible set. Model (e) does not permit the identification of *c*-specific effects, even though *S* is admissible. Model (f) permits the identification using measurement of *Z* though no admissible set exists (U, U_1 and U_2 are unobserved).

C excluded from all admissible sets. An intriguing pattern of heterogeneity is described in Figure 1e. 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 backdoor 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 the subsection on Assessing Covariate-Induced Heterogeneity, equation (1).

Note that model (e) is statistically indistinguishable from model (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.

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.

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[5]:

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

where ETT and ETU are the average effect of treatment on the treated and untreated, respectively,[6] that is:

$$\begin{aligned} ETT &= E(Y_1 - Y_0|X=1), \\ ETU &= E(Y_1 - Y_0|X=0). \end{aligned}$$

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 nonzero only if the two groups respond differently to treatment, thus exhibiting heterogeneity.[7]

Xie et al. called the former type-I bias and the latter type-II bias, whereas Morgan and Winship (2007:46-48) 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 biases, 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.

Separating Fixed-Effect from Variable-Effect Bias

To convince ourselves that baseline and variable-effect biases, as defined earlier, 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 Figure 2 and represents the structural equations:

$$y = \beta x + \gamma z + \delta xz + \varepsilon_1$$
$$x = \alpha z + \varepsilon_2$$
$$z = \varepsilon_3,$$

where the disturbances ε_1, ε_2, and ε_3 are assumed to be mutually independent. Indeed, for variable-effect bias, we obtain[8]:

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

whereas for baseline bias, we have:

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

(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, and it will not be revealed through the difference $ETT - ETU$, unless Z also affects the treatment assignment X.

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

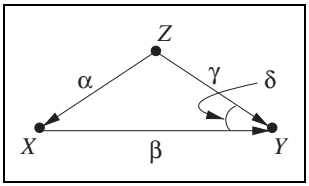

Figure 2. A linear model with interaction, demonstrating baseline and variable-effect biases. The former is proportional to $\gamma\alpha$ and independent of δ; the latter is proportional to $\delta\alpha$ and independent of γ, reflecting effect variability.

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 as follows:

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 following subsections deal separately with each of these cases.

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 identified (Pearl 2009:396-97). 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 $p = P\ (X = 1)$.[9]

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 pretrial and posttrial 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 nonzero $ETT - ETU$, 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, equation (3) implies that whenever we observe the inequality:

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

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 counterintuitive and should be considered a victory of formal counterfactual analysis. The fifth section presents numerical examples of such findings and Online Appendix A provides an example where equation (5) returns equality despite rampant heterogeneity.

Sander Greenland suggested (personal communication, January 24, 2015) that heterogeneity in randomized trials is related to the issue debated by Fisher versus 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, the inequality in equation (5) can be regarded as a testable condition for rejecting Fisher's null hypothesis.

The fifth section and Online Appendix A present models where Neyman's mean null holds, $E(Y_1) = E(Y_0)$, as well as inequality in equation (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 (Ding 2014; Greenland 1991).

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 note 2) the adjustment formula:

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

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 (Shpitser and Pearl 2009), 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 equation (6), save for using $P(z|x')$ as a weighting function, instead of $P(z)$.[10]

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.[11]

When the set Z is large, the estimation of equation (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 to the discussion of identification (Pearl 2009:348-52).

An objection might be raised to classifying the heterogeneity detected by equation (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 the subsection on Special Cases). To witness, equation (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.[12]

Detecting Heterogeneity Through Mediating Instruments

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, which appears to require milder assumptions (Angrist and Pischke 2010; Pearl 2015). 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, Imbens, and Rubin 1996; Brand and Thomas 2013).

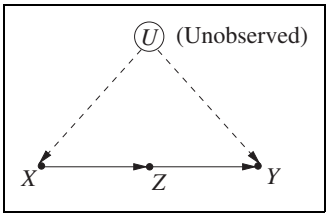

Figure 3. A model in which variable *Z* acts as a mediating instrument for identifying the causal effect of *X* on *Y* in the presence of unknown or unobserved confounders (*U*).

Fortunately, the use of mediating instruments overcomes these limitations and identifies causal effects in nonparametric 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 Figure 3 and assumes the availability of covariates Z that intercept all directed paths from treatment (X) to outcome (Y).[13] Moreover, the graphical theory of *ETT* teaches us that both *ETT* and *ETU* are identifiable in the model of Figure 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, equation (7), save for exchanging x and x'. Moreover, and in contrast to identification by randomized experiment, this estimand remains valid for nonbinary treatments as well.

Accordingly, the estimand for the heterogeneous component of the bias becomes identical to that of equation (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 versus heterogeneous mechanisms on the observed bias without knowing the actual mechanisms involved.

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 percent higher than without training, 0.9 versus 0.8. Let $Z = 0$ stand for the class of uninformed individuals, for whom the chances of hiring after training are 70 percent higher than without training, 0.8 versus 0.1. We will assume that the propensity for enrollment among the informed, q_2, is higher than that among the uninformed, q_1, that is, $q_2 - q_1 = P(X = 1|Z = 1) = P(X = 1|Z = 0) > 0$.

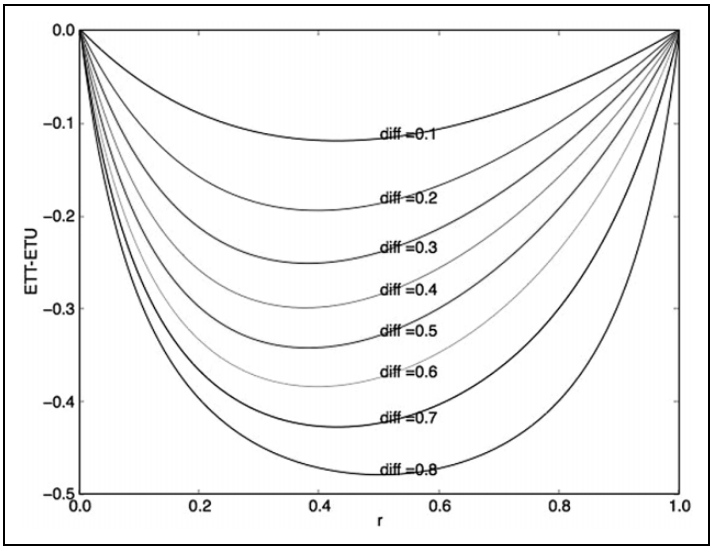

Figure 4. *ETT − ETU* versus *r* for different levels of enrollment disparity, $q_2 - q_1$.

Since we are dealing with a binary treatment, we can assess the magnitude of $ETT - ETU$ using equation (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 Figure 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 $q_2 - q_1$.

In generating these plots, we assume a model similar in structure to the one in Figure 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 heterogeneous individuals. Plots such as those in Figure 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 equation (3)), we can then infer that, if the propensity difference $q_2 - q_1$ is lower than 0.5, the proportion r must lie between 0.20 and 0.62. The larger the difference $q_2 - q_1$, the wider the bounds for r.

Conclusions

This article 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 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 nonbinary 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 Stephen Morgan for calling my attention to the sociological literature on heterogeneity and commenting on earlier versions of the manuscript. Subsequently, this article benefitted from discussions with Felix Elwert and Sander Greenland. I thank Ang Li for generating the plots of Figure 4.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

Notes

1. This example is taken from Pearl (2013).
2. 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. The logic of potential outcomes (Rosenbaum and Rubin 1983) and its equivalence to structural equations where established in (Simon and Rescher 1966; Balke and Pearl 1994ab; Galles and Pearl 1998; Halpern 1998; Pearl 2015).
3. By "admissible," we mean a set C of covariates that satisfy the backdoor criterion (Pearl 1993; Pearl 2009:79-81) in the causal diagram and thus permit 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 backdoor criterion provides a scientific basis and a transparent test for conditional ignorability–type claims, which many researchers entrust to intuition.
4. In practice, the summation over S can be prohibitive, and propensity score weighting can be used over the unit interval $0 \leq PS \leq 1$ (Brand and Xie 2010).
5. This decomposition follows from the consistency rule: $E(Y_1 \mid X = 1) = E(Y \mid X = 1)$, $E(Y_0 \mid X = 0) = E(Y \mid X = 0)$. It was first proposed in sociology by Winship and Morgan (1999:667) 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 coworkers (Heckman 1992; Heckman and Robb 1986).
6. 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.
7. Heckman et al. (2006) called this difference *essential heterogeneity*.
8. These expressions follow directly from the structural definition of counterfactuals (Pearl 2009:98) as defined in equation (12). A complete derivation is given in Online Appendix B.
9. 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 nonbinary treatments, *ETT* is not expressible in terms of $E(Y_0)$ and $E(Y_1)$.

10. This difference accounts for the modified Horvits–Thompson weights required for estimating *ETT* and *ETU* by regression (Morgan and Winship 2015:231).
11. 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.
12. 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 2009:180-82; 2014). Here we witness a causal, not associational relationship that is present in the combined population and is absent in each and every subpopulation.
13. For application of the front-door criterion in the social sciences, see Chalak and White (2012) and Morgan and Winship (2007, 2015).

Supplemental Material

The online appendices are available at http://smr.sagepub.com/supplemental.

References

Angrist, J. D., G. Imbens, and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables (with Comments)." *Journal of the American Statistical Association* 91:444-72.

Angrist, J. D. 1998. "Estimating the Labor Market on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica* 66:249-88.

Angrist, J. D. and A. B. Krueger. 1999. "Handbook of Labor Economics." Pp. 1277-366 in *Causality: Statistical Perspectives and Applications*, 1st ed., vol. 3, edited by O. Ashenfelter and D. Card. Amsterdam, the Netherlands: Elsevier.

Angrist, J. D. and J.-S. Pischke. 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 J. Pearl. 1994a. "Counterfactual Probabilities: Computational Methods, Bounds, and Applications." Pp. 46-54 in *Uncertainty in Artificial Intelligence 10*, edited by R. L. de Mantaras and D. Poole. San Mateo, CA: Morgan Kaufmann.

Balke, A. and J. Pearl. 1994b. "Probabilistic Evaluation of Counterfactual Queries." Pp. 230-37 in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I, edited by B. Hayes-Roth and R. E. Korf. Menlo Park, CA: MIT Press.

Blyth, C. 1972. "On Simpson's Paradox and the Sure-thing Principle." *Journal of the American Statistical Association* 67:364-66.

Brand, J. E. and C. N. Halaby. 2005. "Regression and Matching Estimates of the Effects of Elite College Attendance on Educational and Career Achievement." *Social Science Research* 35:749-70.

Brand, J. E. and J. S. Thomas. 2013. "Causal Effect Heterogeneity." Pp. 189-213 in *Handbook of Causal Analysis for Social Research*, chap. 11, edited by S. L. Morgan. Dordrecht, the Netherlands: Springer.

Brand, J. E. and Y. Xie. 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 H. White. 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 C. Winship. 2010. "Effect Heterogeneity and Bias in Main-effects-only Regression Models." Pp. 327-36 in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Halpern. Milton Keynes, U.K.: College Publications.

Galles, D. and J. Pearl. 1998. "An Axiomatic Characterization of Causal Counterfactuals." *Foundation of Science* 3:151-82.

Greenland, S. 1991. "On the Logical Justification of Conditional Tests for Two-by-two Contingency Tables." *The American Statistician* 45:248-51.

Halpern, J.1998. "Axiomatizing Causal Reasoning." Pp. 202-10 in *Uncertainty in Artificial Intelligence*, edited by G. Cooper and S. Moral. San Francisco, CA: Morgan Kaufmann; *Journal of Artificial Intelligence Research* 12:17-37, 2000.

Heckman, J. 1992. "Randomization and Social Policy Evaluation." Pp. 201-30 in *Evaluations: Welfare and Training Programs*, edited by C. Manski and I. Garfinkle. Cambridge, MA: Harvard University Press.

Heckman, J. and R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156-245 in *Longitudinal Analysis of Labor Market Data*, edited by J. Heckman and B. Singer. New York: Cambridge University Press.

Heckman, J. and R. Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 63-107 in *Drawing Inference from Self Selected Samples*, edited by H. Wainer. New York: Springer-Verlag.

Heckman, J., S. Urzua, and E. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88:389-432.

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

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

Morgan, S. L. and J. J. Todd. 2008. "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects." *Sociological Methodology* 38:231-81.

Pearl, J. 1993. "Comment: Graphical Models, Causality, and Intervention." *Statistical Science* 8:266-69.

Pearl, J. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669-710.

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

Pearl, J. 2013. "The Curse of Free-will and the Paradox of Inevitable Regret." *Journal of Causal Inference* 1:255-257.

Pearl, J. 2014. "Understanding Simpson's Paradox." *The American Statistician* 68: 8-13.

Pearl, J. 2015. "Trygve Haavelmo and the Emergence of Causal Calculus." *Econometric Theory* 31:152-79. Special issue on Haavelmo Centennial.

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Shpitser, I. and J. Pearl. 2006. "Identification of Conditional Interventional Distributions." Pp. 437-44 in *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence*, edited by R. Dechter and T. Richardson. Corvallis, OR: AUAI Press.

Shpitser, I. and J. Pearl. 2009. "Effects of Treatment on the Treated: Identification and Generalization." Pp. 514-21 in *Proceedings of the Twenty-fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*, edited by J. Bilmes and A. NgCorvallis, OR: AUAI Press.

Simon, H. and N. Rescher. 1966. "Cause and Counterfactual." *Philosophy and Science* 33:323-40.

Simpson, E. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society, Series B* 13:238-41.

VanderWeele, T. and J. Robins. 2007. "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology* 18:561-68.

Winship, C. and S. L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659-706.

Xie, Y., J. E. Brand, and B. Jann. 2012. "Estimating Heterogeneous Treatment Effects with Observational Data." *Sociological Methodology* 42:314-47.

Author Biography

Judea Pearl is Chancellor's professor of computer science and statistics at UCLA. He is a graduate of the Technion, Israel, and has joined the faculty of UCLA in 1970, where he currently directs the Cognitive Systems Laboratory and conducts research in artificial intelligence, human cognition and philosophy of science. Pearl has authored three books, Heuristics (1983), Probabilistic Reasoning (1988) and Causality (2000, 2009), winner of the London School of Economics Lakatosh Award. He is a member of the National Academy of Sciences, and a fellow of the cognitive science society and the Association for the Advancement of Artificial Intelligence. In 2012, he won the Technion's Harvey Prize and the ACM Alan Turing Award.