Separated at Birth: Statisticians, Social Scientists, and Causality in Health Services Research

Bryan E. Dowd

Objective. Health services research is a field of study that brings together experts from a wide variety of academic disciplines. It also is a field that places a high priority on empirical analysis. Many of the questions posed by health services researchers involve the effects of treatments, patient and provider characteristics, and policy interventions on outcomes of interest. These are causal questions. Yet many health services researchers have been trained in disciplines that are reluctant to use the language of causality, and the approaches to causal questions are discipline specific, often with little overlap. How did this situation arise? This paper traces the roots of the division and some recent attempts to remedy the situation.

Data Sources and Settings. Existing literature.

Study Design. Review of the literature.

Key Words. Econometrics, biostatistical methods, epidemiology, instrumental variables, observational data/quasi-experiments

COMMON QUESTIONS, DIFFERENT PERSPECTIVES

Determining whether changing the value of one variable actually has a casual effect on another variable certainly is one of the most important questions faced by the human race. In addition to our common, everyday experiences, all of the work done in the natural and social sciences relies on our ability to learn about causal relationships.

Yet the literature on causal inference must be one of the oddest literatures in all of academia. Where else in the modern sciences can one find literature reviews that routinely begin in the 1700s (e.g., Holland 1986); where the past 100 years of research is labeled “a century of denial” (Pearl 1997); where Bertrand Russell is called stupid (Kempthorne 1978); and where pillars

of the field seem oblivious to decades-old developments in their own area of expertise? How is it that the most important question in all of empirical analysis could remain so controversial after being considered for so long a time by so many brilliant minds?

Health services research is an applied field in which analysts from a wide variety of disciplines, including epidemiology, sociology, economics, political science, medicine, law, and operations research, and often trained in the best academic departments in the country, come together to work on problems of immense importance that undeniably involve causal hypotheses. Yet many health services researchers have been trained in disciplines that are reluctant to use causal language, and approaches to causality vary dramatically by discipline.

It is not unusual for a health services research study section (the group of experts who review research proposals and make funding recommendations) to include analysts who maintain that only randomized control trials (RCTs) yield valid causal inference, sitting beside analysts who have never randomized anything to anything. Two analysts debating the virtues of instrumental variables (IV) versus parametric sample selection models might be sitting next to analysts who never have heard of two-stage least squares.

Academic disciplines routinely take different approaches to the same question, but it is troubling when approaches to the same problem are heterogeneous across departments and homogeneous within departments and remain so for decades—suggesting an unhealthy degree of intellectual balkanization within the modern research university. It is one thing to disagree with your colleagues on topics of common interest. It is another thing to have no idea what they are talking about.

Why should the multidisciplinary field of health services research care about controversies over causal inference? In addition to satisfying their intellectual curiosity, health services researchers would observe that the federal government is embarking on a multibillion dollar enterprise known as “comparative effectiveness research” (CER) to compare health outcomes and costs for different types of treatments. CER is pure causal analysis. There already is a vigorous debate about the way that CER should proceed (Worrall 2002; Berwick 2008). Should it be limited to RCTs, or should it include analyses of observational data? If observational analyses are permitted, which specific

Address correspondence to Bryan E. Dowd, Ph.D., Division of Health Policy and Management, School of Public Health, University of Minnesota, PO Box 729 MMC, Minneapolis, MN 55455; e-mail: dowdx001@umn.edu.

methods are preferred? How should the heterogeneous treatment effects inherent in “personalized medicine” be analyzed? These questions need to be resolved. It would be a disservice to patients, health care providers, and taxpayers to have the “experts” disagree at the end of the day about CER’s actual contribution to medical knowledge.

The purpose of this paper is to trace the recent history of controversies in causal inference. This paper will not resolve those controversies, but hopefully it will result in improved communication and shared understanding of the different perspectives.

WHAT HAPPENED, WHEN DID IT HAPPEN, AND WHY DID IT HAPPEN?

The “what” and “when” questions are relatively easy to answer. “Why” is more difficult.

Francis Galton, Karl Pearson, and Correlation

Although many reviews of causality begin with Aristotle and Plato and work their way through David Hume in the 1700s (e.g., Holland 1986), the most recent methodological split in causal inference began in the late 1800s with the work of Francis Galton and Karl Pearson.

Karl Pearson, Francis Galton’s student, credits Galton with the discovery of correlation (Pearson 1920). Pearson was captivated by correlation, but he rejected any notion of causation beyond correlation. For that reason, Judea Pearl refers to Pearson as “causality’s worst adversary” (Pearl 2000, p. 105).

To put Pearson’s objection in context, consider an outcome variable of interest, denoted Y . The value of Y for the i th subject is Y_i . X_i refers to the set of variables that might, in the ordinary language of everyday experience, have a causal influence on the value that Y assumes. The X variables could include the primary causal variable or mediators or surrogates by which the causal effect is transmitted from a more “fundamental” cause to Y . An X of particular interest is one that designates membership in the treatment group (T) versus the control group. T can be either a discrete or continuous variable. Although the distinction between a continuous versus discrete T has important implications for estimation, the discussion of the issues in this section applies to a one unit change in either type of treatment. Also, there could be many different treatments and many types of control groups, but for simplicity, the examples in this paper are limited to one of each. Although the language of causality can be

controversial, it is important to remember that the entire purpose of forming and talking about treatment and control groups in the first place is that one thinks it is possible that the treatment might have a causal effect on the outcome of interest.

I state the fundamental causal question as, “What is the change in the expected value of Y , or the probability that Y assumes a particular value, when the value of T is changed, by external means, by one unit?” The emphasis on *change* separates questions of causality from questions of association. By “external means” I mean that the change in T is induced by some mechanism that is uncorrelated with the unobserved factors that affect Y . Randomization assigns research subjects to the treatment versus control group, but the act of randomization, per se, is assumed to have no direct effect on the outcome variable of interest. Similarly, the distance from a person’s place of residence to a hospital that offers one type of treatment for acute myocardial infarction (AMI) versus another could affect the treatment an AMI patient receives, but that distance, per se, plausibly has no direct effect on the patient’s immediate health outcome resulting from treatment at the hospital.

This version of the causal question reflects specific positions on a host of interesting and controversial questions. For example, references to the causal effect of X on the *expected* value of Y , or the *probability* that Y takes on a specific value are a relatively recent development in the literature (Suppes 1970). Cartwright (2007) objects to the assumption that the values of other X variables can be held constant in observational studies when the value of T changes.

In some of the discussion, it will be helpful to refer to a linear model that takes the form¹:

$$Y_i = X_i\beta + T_i\beta_T + u_i \quad (1)$$

where u is unobserved error. The β ’s are coefficients. The model is written in linear form for simplicity, but all the discussion in this paper can be generalized to models that are nonlinear in either the variables or the parameters, even though different estimation approaches are required.

Equation (1) has no causal interpretation, per se, as long as one is restricted to information contained in the numeric values of Y , X , and T . The numeric values of Y , X , and T cannot reveal which way the causal arrows go in Figure 1. The direction of the arrows relies on information beyond the data in hand, for example, knowing that X assumes its value in a time period before the time period during which Y assumes its value. That knowledge does not establish a causal relationship between X and Y because many events could precede Y and still be causally unrelated to Y , or X and Y could be the result of

Figure 1: An Observed Confounder (W)

some common cause W , but the knowledge that X precedes Y does at least rule out $Y \rightarrow X$.

Pearson appears to have taken a conservative approach to allowing such extraneous information to inform the assessment of causal effects. However, as discussed later in the paper, if the restriction on information beyond the numeric values of Y , X , and T is enforced strictly, causal inference will be the least of the analyst's worries. The controversies in the causal literature center on precisely *what* information beyond the numeric values of Y , X , and T the analyst will be allowed to consider.

Now, suppose that in addition to the variables T and Y we add a confounder variable W that has a causal effect on both T and Y , as shown in Figure 1. If W is unobserved, the resulting estimate of the causal effect of T on Y (β_T) will be biased. In that case, some analysts might refer to W as an unobserved confounder, while others would say that T and Y are “spuriously correlated.” Econometricians would refer to the problem as omitted variable bias.

If Figure 1 is true (the crucial assumption) and W is observed, an unbiased estimate of β_T is the partial correlation of T and Y controlling for W in standardized data (i.e., $\frac{X_i - \bar{X}}{\sigma_X}$ for all variables) and is computed by regressing T on W and Y on W in separate regressions. The correlation of the residuals from those two equations is β_T .

From our vantage point, it seems obvious that the ability to calculate the correlation of T and Y controlling for W , thereby reducing (though not necessarily eliminating) the bias in β_T , would have been viewed as a major improvement in causal analysis. Apparently Pearson did not see it that way. There are several possibly explanations. First, Figure 1 is filled with untestable assumptions if one is restricted to information contained in the numeric values of Y , X , and T . Second, the computational difficulty of controlling for multiple

W variables in regression equations in the early 20th century posed a formidable practical limitation on causal analysis.²

Pearson's reticence regarding information beyond the numeric values of Y , X , and T may have been due in part to the limited number of ways in which the value of T could arise when he published his history of correlation in 1920. Six years later, Ronald Fisher discovered the randomized trial.

Ronald Fisher and Philip Wright

Two major discoveries in the 1920s defined the alternative empirical approaches to causal analysis that, over time, split the field of statistics: the discovery of random assignment by Ronald Fisher (1926) and the discovery of IV by Philip Wright (1928). The split is shown in Figure 2.

The key to the RCT was the analyst's ability to manipulate the causal variable of interest (T). That approach was quite feasible in Fisher-type problems, for example, effect of different fertilizers (T) on crop yields (Y). The great attraction of randomization was its claim to render all unobserved confounding variables causally irrelevant, a claim questioned by Urbach (1985) and others.

The key to IV estimation (Appendix SA1) was the analyst's ability to identify an IV that, like randomization, affected the assignment of individuals to the treatment group versus the control group, but was uncorrelated with the unobserved factors that affect the outcome. The great attraction of IV estimation was that it could be applied to problems where random assignment to potentially endogenous³ explanatory variables was difficult or impossible.

Figure 3 shows a causal model in which unobserved variables (v) that affect or are affected by⁴ assignment to the treatment group (T) are correlated with unobserved variables (u) that affect the outcome variable of interest (Y), resulting in correlation of T and u , and thus biased estimates of β_T .⁵ The variable Z represents the variable that affects assignment of subjects to the treatment versus control group. Z could be either randomization or an IV. The important characteristic of Z is that it is uncorrelated with u . That characteristic is controversial in the case of IV because it is not directly testable using only the data represented by the variables in Figure 3. That point is discussed in greater detail later in the paper.

Fisher's discovery was well-received and his experimental approach to causal inference came to dominate the field of statistics, in part, because random assignment helped to link the relatively new field of statistics to the prestigious natural sciences. The statistician who could calculate the comb-

Figure 2: Historical Overview of Developments in Causal Inference

Note. The diagram is difficult to draw because it combines chronology (correlation precedes IV) with typologies. Also, the typologies might be controversial. For example, some analyst might view regression discontinuity as an experimental method. The term “longitudinal” data could apply either to observational data or to data from an RCT if there were multiple preintervention or postintervention observations on subjects.

inatorics for a Latin Square design and lead an RCT became an integral part of the natural science research team.

Wright’s discovery of IV, on the other hand, was ignored for approximately 20 years before being rediscovered by the Cowles Commission after World War II (Stock and Trebbi 2003, p. 182). Even when econometricians adopted Wright’s approach, they largely eschewed explicit causal language. Pearl (1997) refers to the 20th century as a “century of denial” that valid causal inference could be drawn from observational data and laments “an alarming tendency among economists and social scientists to view a structural equation⁶ as an algebraic object that carries functional and statistical assumptions but is void of causal content.” He wonders, “. . . what has happened to (structural

Figure 3: Causal Diagram of an Instrumental Variable (Z)

equation modeling) SEM over the past 50 years, and why the basic (and still valid) teachings of Wright, Haavelmo, Marschak, Koopmans, and Simon have been forgotten” (Pearl 2000, pp. 135–7). Pearl (2000, p. 138) suggests that causal language was abandoned by SEM proponents in an attempt to mollify statisticians, “the arbiters of respectability.” Even today, it is surprising how much print in econometrics textbooks is devoted to the task of obtaining estimators with desirable large and small sample properties, for example, unbiasedness, consistency and efficiency, and how little is devoted to explaining exactly what one has estimated in an unbiased, consistent, and efficient manner.

Bertrand Russell and Physical Laws

Bertrand Russell was one of the towering figures in the fields of mathematics and philosophy in the early to mid 20th century. Russell’s views on causality had a profound influence on the developing field of statistics. One of his most famous quotes is the following:

All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word “cause” never occurs . . . The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm. (Russell 1913)

If causal relationships would not conform to the physical laws, Russell was prepared to abandon causality. Russell maintained that causal relationships needed to have three attributes which he thought characterized all physical laws: (1) causal symmetry; (2) irrelevance of time ordering (time invariance); and (3) determinism (no stochastic processes involved).⁷

Later developments in quantum physics proved Russell's assertions wrong,⁸ but Kempthorne (1978) cites a more fundamental problem with Russell's analysis: failure to recognize the importance of experimentation:

... I, a lowly statistician, am compelled to regard Russell as being very stupid in this connection, for the reason that he did not, it would seem, give the slightest recognition of the idea of experimentation. (Kempthorne 1978, p. 8)

Nonetheless, Russell's analogy to physical laws was an influential force in the development of statistics. Philip Wright anticipated that his estimation approach to omitted confounders would face enthusiastic opposition. He was careful to state that: "Estimates of [demand and supply]⁹ elasticities may be made, but any hope of obtaining numerical values comparable with results to be obtained in physical science must be abandoned." The desire to maintain the link between statistics and the natural sciences survived despite Russell's errors and the irony that some of the most stunning achievements in physics in the early 20th century were based on observational data and conjectures for which no empirical test was possible at the time (e.g., Einstein's conjecture that identical clocks placed at the equator and either of the poles would keep different time).

The Split Solidifies

Over time, the different empirical approaches to causal analyses pioneered by Fisher and Wright solidified into the "experimental" versus "observational" approaches to empirical investigations of causality. But the mere development of two different approaches to causal analysis, by itself, cannot explain the deep and long-lasting intellectual divide between statisticians and structural equation modelers (SEMs) that came to include sociologists, econometricians, and other representatives largely from the social sciences. As Robert Moffitt (1996, p. 462) notes:

IV (instrumental variables) is widely regarded by economists as one of the most versatile and flexible techniques, applicable in an enormous number of disparate applications. Yet it is scarcely used or discussed by statisticians, who often do not see the point of it all.

There are several factors that contributed to the depth and endurance of the split. The first is that statisticians and SEMs often were working on different types of problems. The statisticians' problems often fit neatly into the $T \rightarrow Y$ framework of Figure 1. Wright's model, based on the intersection of simultaneously but separately determined supply and demand curves, was

very different from the standard problem of omitted variable bias, even the IV version shown in Figure 3.

Second, statisticians often were working on a very special subset of the problems in Figure 1 where it was possible for the analyst to manipulate T , for example, through random assignment. The compelling appeal of randomization left statisticians suspicious of empirical investigations of causality where such manipulation was impossible. Holland (1986, pp. 954, 959) describes the statistician's position:

Put as bluntly and contentiously as possible, in this article I take the position that causes are only those things that could, in principle, be treatments in experiments.

Donald Rubin and I once made up the motto

NO CAUSATION WITHOUT MANIPULATION

to emphasize the importance of this restriction.

This remarkable restriction eliminates race and gender as the cause of being denied a promotion or receiving inferior health care (Holland 1986, p. 946; Shadish, Cook, and Campbell 2002, pp. 7–8)—often a surprise to health services researchers engaged in disparities research. Holland elaborates:

As an example, the schooling a student receives can be a cause, in our sense, of the student's performance on a test, whereas the student's race or gender cannot. (Holland 1986, p. 946)

Variables like race or gender are termed "attributes" of subjects because they are not amenable to manipulation, and attributes cannot be causes.¹⁰ Statisticians might agree, however, that racial *discrimination* could cause people to be denied employment, proper health care, and so on, because racial discrimination, at least in theory, could be modified by the analyst.

The quote also is notable in light of the fact that at the time Holland's article was written Donald Rubin and Paul Rosenbaum were collaborating on the development of propensity score analysis of observational data (Rosenbaum and Rubin 1983, 1984), following Rubin's (1974) endorsement of "carefully controlled" analyses of nonrandomized data to estimate causal effects as a "reasonable and necessary procedure in many cases" (p. 688).

Objection to "Untestable Assumptions"

Although the split between statisticians and SEMs may have begun with different types of problems that were amenable to different empirical

approaches, the difference often is formalized by statisticians into an objection regarding the untestable assumption of IV estimation, specifically that the variable Z in Figure 3 affects T but is uncorrelated with u . There are several points to make about objections to untestable assumptions.

First, what information should the analyst be allowed to use in tests of a model's assumptions? Should it be only the information contained in the numeric values of Y , X , and T ? If so, the analyst is in deep trouble. Staring at a sheet of paper containing the values of Y , X , and T , how can the analyst know for sure that the variable in the first column actually is "diastolic blood pressure" rather than the last 100 winning numbers in the state lottery? How does the analyst know that the blood pressure variable was collected before the variable indicating whether the patient experienced a stroke? How does the analyst know that the data truly were generated by an RCT rather than a clever matching strategy? Causal inference in any setting, whether it is daily application of common sense or sophisticated data analysis, requires some assumptions about the process that generated the data and it often will be impossible to test those assumptions using only the single dataset in hand.

Second, statistical analyses are filled with untestable assumptions. What is the empirical test for the existence of the hypothetical repeated samples that underlie frequentist statistics? One can draw repeated samples from a known population, but when one has only a sample of data, generalizing to any conceptual population becomes a matter of speculation. What is the empirical test that any single sample, even one drawn from a population with a previously estimated mean, is not in fact a member of a "relevant subset" of the population that has a different mean (Johnstone 1989)?¹¹ What is the empirical test that randomization was successful with respect to *unobserved* confounders? What is the empirical test that the Bayesian's priors (whether informative or not) are, in any sense, the best approximation to objective rather than subjective reality? In theory, these untestable assumptions could be prioritized, but until they are, broad objections to untestable assumptions minimally lack precision.

Third, discomfort with untestable assumptions can bleed over into a general aversion regarding models of relationships between unobserved variables. Holland (1986, p. 946) reminds us that statistics is concerned with measurement, and it is difficult to imagine measuring what one cannot observe.¹² Statisticians seem comfortable with models of heteroscedasticity and autocorrelation, for example, that incorporate relationships among unobserved variables, and indeed those assumptions can be subjected to specification tests. But how is the assumption that the error terms for two people in a regression are correlated (perhaps because they belong to the same family) any more testable than the

assumption that u and v are in Figure 3 are correlated (the correlation produced by the Heckman–Lee sample selection model)?

Fourth, any assertion that attempts to limit the domain of “legitimate” knowledge is open to a self-referential critique. If an analyst asserts that legitimate causal knowledge can be obtained only from research designs free of untestable assertions, we might ask the analyst to prove that assertion using only information from research designs free of untestable assertions. This problem in logic may seem esoteric to health services researchers, but it is similar to the self-referential critique that led to the demise of logical positivism (the assertion that the only valid data are data obtained from our senses and subject to empirical verification) in philosophy departments in the mid 20th century.

The objection to untestable assumptions needs to be carefully delineated. There can be no general objection to all information beyond that contained in the numeric values of the variables, and once the door is open to additional information then the “untestable” assumption of IV are likely to appear less forbidding.

Here is an example. In studies that compare the health outcomes resulting from two different treatments, a popular IV (Z) in health services research is the distance from the subject’s place of residence to a health care facility that offers one type of treatment versus another. A famous example is McClellan, McNeil, and Newhouse (1994) study of the effect of more intensive treatments on mortality in patients with acute myocardial infarction (AMI). The authors’ assumption was that AMI patients did not choose their place of residence to be closer to a hospital offering one type of treatment for AMI patients versus another. That assumption is testable in two ways. First, one simply could ask the patients how they chose their place of residence. Second, one could agree with the authors that unobserved measures of health status were the variables most likely to represent unobserved confounders in the analysis. Then one could do what the authors did and check to see whether the distance variable was correlated with *observed* measures of health status. Since it was not, the authors concluded that it was unlikely that distance was correlated with *unobserved* measures of health status. The same comparison of observed variables in the treatment and control groups is used to test that randomization was successful.

NO FREE LUNCH

All of the research designs and estimation methods shown in Figure 2 have advantages and disadvantages. All rely on assumptions, many of which are

untestable, especially if one is allowed to use only the abstract numbers in the current data file. All have the ability to produce valid causal inference when the assumptions plausibly are met. Pearl's (2000) directed acyclic graphs provide a framework for identifying the causal assumptions that are necessary to draw valid causal inference from hierarchical models (i.e., models in which the causal arrows all go in one direction). The following comparison of RCTs and IV estimation could be extended to the other observational data methods discussed later in the paper.

There are many disadvantages to IV estimation, in addition to the difficulty of testing the lack of correlation of Z and u . IV estimation is known to perform poorly in the presence of weak instruments (Murray 2006).

In some situations, the estimated treatment effect easily could be heterogeneous (varying from one subject to another). But the estimated treatment effect also might vary from one type of Z to another. For example, one might not expect the same response to a smoking cessation program if the subjects enrolled in it because (a) they were encouraged to do so by a friend; (b) they were encouraged to do so by their employer; or (c) they were paid to do so.

In other situations, the IV Z may be an important determinant of assignment to the treatment versus control group for only a portion of the subjects. Some subjects may have chosen the treatment or control group regardless of the value of the instrument. Patients for whom the instrument is an important determinant of choice of the treatment versus control group are said to be "on the margin" with respect to the instrument.

If the effect of the treatment is homogeneous across subjects, then the treatment effects estimated for patients on the margin will be identical to those for the rest of the sample. If the treatment effect is different for different subjects (heterogeneous), then it is important to ask, "Who is the marginal subject?" (Harris and Remler 1998). Are subjects for whom the instrument works well the same subjects for whom the treatment has a significant effect? Models with heterogeneous treatment effects are more difficult to estimate in any research design, including RCTs, and require samples large enough to estimate the different treatment effects in each subsample.

Although RCTs are vulnerable to attrition bias, compliance problems and cross-contamination of the treatment and control groups, an important advantage of RCTs is that randomization affects assignment equally for *all* the subjects in the experiment. Nonetheless, heterogeneous treatment effects still are a problem. In extreme cases of heterogeneous treatment effects, an RCT cannot distinguish between a treatment that has no effect on the outcome and one that either kills the patient or saves her life. The problem is that

unobserved variation in the patient's medical condition can act as a moderator of the effect of the treatment on the outcome of interest. An example, adapted from Pearl (2000), is provided in Appendix SA2. Heckman (2008) also notes that RCTs do not permit study of subjects' choice of the treatment versus control group, which could have a dramatic effect on the average treatment effect in real-world applications if subjects who self-select into the treatment group have a different response than those who were randomized to the treatment.

In RCTs it is difficult to ensure close correspondence of either the intervention or the subjects as specified in the RCT to real-world application. Internal validity of a study is improved by careful specification of both the intervention and the eligible subjects, but narrow specification of either reduces the study's external validity, that is, its generalizability.

Critics also cite the expense of RCTs, and the ethical paradox that the cost entails. Unless one has good reason to think that the treatment may be effective, the cost of an RCT is difficult to justify. However, if one has good reason to think that the treatment is effective, it is difficult to justify withholding it from the control group.

Particularly in the small samples that are common in clinical trials, randomization can fail to achieve balance on observed covariates. As Urbach (1985) notes, Fisher's original rationale for randomization was to justify the statistical significance of the test of the null hypothesis (no difference in outcomes in the two groups). But suppose that following randomization, the treatment and control groups are unbalanced on some observable variables. What should the analyst do? The analyst might rerandomize patients to the treatment and control groups and test again for differences in observed variables, but Urbach notes that:

... if randomization purports to underwrite the significance test, then picking and choosing among the distributions thrown up by chance is quite illicit, for this undermines that test. But if we do not allow this kind of post-randomization selection, we might end up with test groups that differ substantially in their relevant characteristics, in which case we shall be assured of reaching a false conclusion. (p. 260)

Alternatively the analyst might use multivariate regression (or propensity scores) to control for differences in observed covariates. But if one is prepared to rely on statistical controls to correct imbalance, why bother with randomization in the first place? The hope is that randomization will balance *unobserved* confounders in the treatment and control groups, but how justified is that hope if randomization has failed to achieve balance on *observed* confounders?

A valid but expensive response to unbalanced groups is to continue with the experiment and then repeat the experiment many times, compiling the results into an unbiased estimate of the average treatment effect computed across all the experiments.¹³ Deaton (2009) discusses other potential difficulties with RCTs.

Additional Observational Data Methods

The impracticality of RCTs for many questions has led to the development of a variety of analytic approaches for observational data in addition to IV. The sample selection model of Heckman (1974) and Lee (1976) jointly estimates an equation that represents endogenous choice of the treatment (e.g., T estimated as a function of Z , additional covariates, and v) along with an outcome equation (e.g., Y estimated as a function of T , additional covariates and u). The correlation of u and v is estimated as one of the model's parameters. The data requirements for the IV and sample selection models are the same,¹⁴ and the sample selection model typically is estimated either by a two-step or a maximum likelihood estimator (MLE) method. The MLE has been criticized for its assumption that the error terms u and v have a bivariate normal distribution, but the two-step estimator can be derived by assuming only that v and u have a linear regression relationship (Olsen 1980). Lee (1983) showed that the MLE can be generalized to *any* distribution of v and u .

One popular choice for an instrument is a “natural” experiment. A natural experiment is an exogenous shock to a system that affects the causal variable of interest, but it is uncorrelated with the error term in the (outcome) equation of interest. A change in tax policy that occurs in one state but not in another might alter the relative price of a good or service in ways that facilitate study of the effects of prices on consumption. Of course, the obvious question is, “Why did the change in tax policy occur in one state but not another?” The analyst must argue that the (unobserved) factors that caused the change in policy have no *direct* causal connection to the dependent variable except through the price of the taxed commodity.

When data are available on the same subjects before and after the intervention (e.g., the occurrence of the natural experiment's shock) difference-in-difference (DID) models can be used to compare changes in the dependent variable in the preintervention and postintervention time periods and contrast those changes for the treatment and control groups. DID models based on only one observation per subject in the preintervention and postintervention time periods cannot capture differences in time trends in either period. When

multiple observations are available on the same subject in the preintervention and postintervention time periods, panel data methods (referred to in some literatures as “interrupted time series”) can be used to estimate trend lines in the preintervention and postintervention periods. In observational data, fixed effects for the research subjects often are used to control for the effects of time-invariant characteristics of the subject that could represent unobserved confounders.

Regression discontinuity models are applicable when a break point in a continuous assignment variable divides the sample into the treatment and control groups (Cook 2008). An example is school children who either are or are not given remedial help based on a break point in a baseline test score. Subjects just to either side of the break point are assumed to be similar. Causal inference from regression discontinuity models can be sensitive to accurate modeling of the often nonlinear relationship between the assignment variable and outcome variable.

CONCLUSION

Health services research has become an intellectual crossroads for RCTs and SEMs because many research questions are amenable to both RCT and SEM approaches. Contentious situations arise when the advantages and disadvantages of both approaches must be weighed in an environment of limited funding for research.

A number of authors have made valiant attempts to bridge the disciplinary divides in causal analyses. Judea Pearl’s work is a shining example. Sander Greenland (2000) introduced IV to epidemiologists where it appears to have found a receptive audience (see, e.g., Hernan and Robins 2006; Martens et al. 2006; Rassen et al. 2009a, b). Angrist, Imbens, and Rubin (1996) made a similar effort with statisticians. Heckman has been writing specifically about causality for over a decade. He recently compared econometric approaches with those more familiar to statisticians (Heckman 2008) and sociologists (Heckman 2006) and attempted to find common ground between structural and program evaluation approaches to policy analysis (Heckman 2010).¹⁵ Bridge building can be a risky endeavor, however. The exchanges between Heckman (1996) and Angrist, Imbens, and Rubin (1996) and the contributors to the June 2000 issue of the *Journal of the American Statistical Association* reveal the depth of the divides. Not all the disagreements are cross-disciplinary. A collection of articles in the Spring 2010 issue of the *Journal of Economic Perspectives* reveals deep divides

among econometricians focused solely on program evaluation problems. An encouraging sign is the increased frequency of articles that compare results from different methods (Earle et al. 2001; Stukel et al. 2007).

Holland (1986, p. 947) defines the fundamental problem of causal inference as the impossibility of observing pure counterfactuals, for example, the outcome variable measured for the same subject at the same point in time under both treatment and control conditions.¹⁶ I prefer Frank Knight's statement of the more general problem of acquiring knowledge, quoted by Heckman, Lalonde, and Smith (1999, p. 20):

The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of knowledge depends on the future being like the past. (Knight 1921, p. 313)

How do we decide what parts of the past are stable and what parts can change in ways that do not threaten our stability assumptions? SEMs generally assume that the structural form equations and their coefficients will remain "like the past" when the values of the explanatory variables, including the variable representing the treatment, are changed to new values that are "not like the past."¹⁷ Advocates of RCTs generally assume that the structure of the experiment, for example, the method of drawing the sample, the assignment of subjects to the treatment and control group, the treatment itself, and the measurement of outcomes, are both replicable and representative of future applications in the "real world." What is the justification for those assumptions? Here it is worth quoting Michael Polanyi (1968, p. 38) at length:

If the orthodox theory of scientific explanation is misleading, the treatment of empirical generalization is equally so. Without going into detail, I may point out here three major errors which have resulted from the attempt to define empirical validity by strict criteria. First, since no formal procedure could be found for having a good idea from which to start on an enquiry, philosophers virtually abandoned the attempt to understand how this is done. Second, having arrived at the conclusion that no formal rule of inference can establish a valid empirical generalization, it was denied that any such generalization could be derived from experimental data—while ignoring the fact that valid generalizations are commonly arrived at by empirical enquiries based on *informal* procedures. Third, it was claimed that a hypothesis is strictly refutable by a single piece of conflicting evidence—an illusory claim, since one cannot formally identify a contradictory piece of evidence.

Polanyi, like Thomas Kuhn (1962), is trying to dispel our romantic notions regarding scientific knowledge. All approaches to knowledge acquisition are

based on an assumed degree of system stability. Polanyi (1962) demonstrates the importance of “tacit knowledge” regarding those assumptions. Reaching agreement on the most plausible set of stability assumptions in any particular application can be difficult, but any attempt to place a priori restrictions on the ways in which we are allowed to learn about anything (like the effectiveness of medical treatments) ultimately will have to confront the vast magnitude and complexity of ways in which we *actually* learn about things.

Human beings, including health services researchers, will continue to draw valid causal inferences from observational data regardless of how many analysts confidently assure them that it is impossible. A productive, but challenging area of causal research is to understand better how human beings manage to draw valid causal inference so quickly and in general, so reliably from observational data. Improved understanding would produce both better analysis of observational data and better experiments.

The challenge for health services research and the health care system in general is to contemplate the physician’s decision problem as she sits across the table from her patient. On what evidence will her treatment decisions be based? A similar case she treated 5 years ago? Results from an RCT only? What if there are not any RCT results or the RCT involved a substantially different form of the treatment applied to patients substantially different from the one sitting across the table? What if the results came from an observational study, but the conditions required for the estimation approach were not fully satisfied?

Judea Pearl (2000) opens his book on causality with a quote from Albert Einstein in 1953:

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

Logical systems and systematic experiments unquestionably have been essential tools of Western scientific advancement, but they are not the foundation of that advancement. The axiomatic distinction that separated the West from the rest was the confident belief that reality was not capricious, but instead was orderly and reliable. It was that belief that made it plausible to imagine that logic and experiments, law-like behavior and correspondence of the physical world to mathematics could be useful tools in the exploration of physical reality—even when reality turned out to have a significant stochastic component at the quantum level. Other civilizations saw the same data

as the West, but they interpreted the data differently because they began with different assumptions. It is the assumptions regarding the data-generating process that matter, and those assumptions always come from beyond the data themselves.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The author would like to acknowledge the helpful comments of two referees, and Paul Glewwe and Kirk Allison, as well as participants in a series of seminars on causality sponsored by the Division of Health Policy and Management, School of Public Health, University of Minnesota during April 2009.

Disclosures: None.

Disclaimers: None.

NOTES

1. This way of writing the causal equation sidesteps the controversial question of whether two separate outcomes Y_i^T and Y_i^C representing “potential” outcomes exist for the same person if s/he had been in the treatment versus the control group, respectively (Dawid 2000).
2. The ordinary least squares estimator of linear regression coefficients is $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$, where $(X'X)^{-1}$ is the inverse of the cross-products matrix of the explanatory variables. Although OLS regression was discovered by Legendre in 1805 (Freedman 1999) and matrix inversion was formalized by Cayley (1854) in the mid-1800s, inverting a 10×10 matrix, for example, remained a formidable task, undertaken by a team of human beings rather than computers, even through World War II.
3. Throughout the paper, I use the term “endogenous” to refer to explanatory variables that potentially are correlated with the error term in a regression format. Two important sources of endogeneity are reverse causality and omitted variables, sometimes referred to as unobserved confounders or sources of spurious correlation (Dowd and Town 2002).
4. For example, physicians might give extra attention to patients in the treatment group in ways not intended as part of the trial. The purpose of blinding physicians to the patient’s treatment status is to eliminate that possibility.
5. This is *not* a diagram of the same problem (supply and demand) that Philip Wright was trying to solve, but it is a diagram of the problem that inspired Ronald Fisher’s discovery of randomized trials. As discussed later in the paper, this discrepancy might account for some of the early failure to connect the two approaches.

6. A structural equation is an equation taken directly from the causal model. For example, *linear* structural equations corresponding to Figure 3 would be $Y = X\beta + T\beta_T + u$ and $T = Z\gamma + v$. Heckman (1999, p. 9) states that the terms “structural” and “causal” models are synonymous and Pearl’s definition (2000, p. 160) is consistent with that interpretation. The term “structural equation model” is used in the literature on path coefficients which can be traced to Sewell Wright (1918). The later path analysis literature contains explicit causal language. Duncan (1966) provides a good discussion.
7. As Rosenberg (1989) concludes in his review of Russell’s writings:

What Russell believed modern science to have shown is that there is no such thing as causal directionality. (Rosenberg 1989, p. 341)

Second:

Time must not enter explicitly into our formulae. All mechanical laws exhibit acceleration as a function of configuration, not of configuration and time jointly; and this principle of the irrelevance of time may be extended to all scientific laws. In fact we may interpret the “uniformity of nature” as meaning just this, that no scientific law involves the time as an argument, unless of course it is given in an integral form, in which case *lapse* of time, though not absolute time, may appear in our formula. (Russell 1913)

And finally:

The essential function which causality has been supposed to perform is the possibility of inferring the future from the past, or more generally, events at any time from events at certain preassigned times. Any system in which such inference is possible may be called a “deterministic” system . . . (Russell 1913)

8. Steiner (1986) and Rosenberg (1989) come to Russell’s defense, noting the time period of his writing. Readable counter-examples to Russell’s characterization of physical laws can be found in two extraordinary best-sellers by Oxford physicist and mathematician Roger Penrose: *The Emperor’s New Mind* (1989), and *The Road to Reality* (2005). For example, imagining one type of experiment running backwards would require X-ray photons to “jump out of the floor” of the laboratory rather than being produced by the photon emitter (2005, p. 821).
9. The original reads “Estimates of *their* elasticities . . .” (Wright 1928 p. 304).
10. Holland’s view of “attributes” appears to be quite broad and includes “scholastic achievement” in the context of a study by Saris and Stronkhorst (1984).
11. The problem of relevant subsets presents a challenge for RCTs. If there are no relevant subsets, then randomization is unnecessary; and if there are relevant subsets, then randomization cannot guarantee that a particular subject does not belong to one of them.
12. The preference for the observed over the unobserved led to a dispute over potential outcomes (see Dawid 2000, and respondents to his article in the June 2000 issue of *The Journal of the American Statistical Association*).

13. This is not the same point as saying that “no causal inference should be drawn from a single study.” The point here is that the estimate from the single study is known to be biased.
14. Although the selection effect in the Heckman–Lee model theoretically is identified by the nonlinearity of the first stage sample selection equation (often probit), Manning, Duan, and Rogers (1987) showed that the performance of the model is improved dramatically by the inclusion of variables that affect only the sample selection process and not the outcome in the equation of interest. Monfardini and Radice (2008) report similar results for the bivariate probit model.
15. Heckman (2010) uses the term “structural” to refer to “. . . parametric, explicitly formulated, empirical economic models,” while “program evaluation” refers to “. . . ‘effects’ defined by experiments or surrogates for experiments”
16. Note that if the nonexistence of strict counterfactuals were a barrier to valid causal inference, that barrier would apply to randomized trials, as well as nonexperimental data.
17. An important exception is the Lucas (1976) critique in economics.

REFERENCES

- Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91 (434): 444–55.
- Berwick, D. 2008. “The Science of Improvement.” *Journal of the American Medical Association* 299 (10): 1182–4.
- Cartwright, N. 2007. “Counterfactuals in Economics.” In *Causation and Explanation*, edited by J. K. Campbell, M. O’Rourke, H. S. Silverstein, and H. Silverstein, pp. 191–216. Cambridge, MA: MIT Press.
- Cayley, A. 1854. “On the Theory of Groups as Depending on the Symbolic Equation $\theta^n = 1$, 7:4 *Philosophical Magazine* 7 (42): 40–47. Also ‘A Memoir on the Theory of Matrices.’” *Philosophical Transactions of the Royal Society of London* (1858): 17–37.
- Cook, T. D. 2008. “‘Waiting for Life to Arrive’: A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics.” *Journal of Econometrics* 142 (2): 636–54.
- Dawid, A. P. 2000. “Causal Analysis without Counterfactuals.” *Journal of the American Statistician* 95 (450): 407–24.
- Deaton, A. S. 2009. *Instruments of Development: Randomization in the Tropics and the Search for the Elusive Keys to Economic Development*. National Bureau of Economic Research, Working Paper 14690 [accessed on October 22, 2010]. Available at <http://www.nber.org/papers/w14690>
- Dowd, B., and R. Town. 2002. “Does X Really Cause Y? Academy Health: Changes in Health Care Financing and Organization” [accessed on October 22, 2010]. Available at <http://www.hcfo.net/pdf/xy.pdf>
- Duncan, O. D. 1966. “Path Analysis: Sociological Examples.” *American Journal of Sociology* 72: 1–16.

- Earle, C. C., J. S. Tsai, R. D. Gelber, M. C. Weinstein, P. J. Neumann, and J. C. Weeks. 2001. "Effectiveness of Chemotherapy for Advanced Lung Cancer in the Elderly: Instrumental Variable and Propensity Analysis." *Journal of Clinical Oncology* 19 (4): 1064–70.
- Fisher, R. 1926. "The Arrangement of Field Experiments." *Journal of Ministry of Agriculture* 33: 500–13.
- Freedman, D. 1999. "Some Remarks on the History of Statistics." *Statistical Science* 14 (3): 243–58.
- Greenland, S. 2000. "An Introduction to Instrumental Variables for Epidemiologists." *International Journal of Epidemiology* 29: 722–9.
- Harris, K. M., and D. K. Remler. 1998. "Who is the Marginal Patient? Understanding Instrumental Variables Estimates of Treatment Effects." *Health Services Research* 33 (5, part 1): 1337–60.
- Heckman, J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42 (4): 679–94.
- . 1996. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 91 (434): 459–62.
- . 1999. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." Working Paper 7333. Cambridge, MA: National Bureau of Economic Research [accessed on October 22, 2010]. Available at <http://www.nber.org/papers/w7333>
- . 2006. "The Scientific Model of Causality." *Sociological Methodology* 35 (1): 1–98.
- . 2008. "Econometric Causality." *International Statistical Review* 76 (1): 1–27.
- . 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48 (2): 356–98.
- Heckman, J., R. Lalonde, and J. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics: Vol. 3, Chapter 31*, edited by O. Ashenfelter and D. Card, pp. 1865–2097. Amsterdam: North Holland.
- Hernan, M. A., and J. M. Robins. 2006. "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology* 17 (4): 360–72.
- Holland, P. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.
- Johnstone, D. J. 1989. "On the Necessity of Random Sampling." *British Journal for the Philosophy of Science* 40 (4): 443–57.
- Kemphorne, O. 1978. "A Biometrics Invited Paper: Logical, Epistemological and Statistical Aspects of Nature-Nurture Data Interpretation." *Biometrics* 34 (1): 1–23.
- Knight, F. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin Company.
- Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
- Lee, L-F. 1976. "Estimation of Limited Dependent Variables by Two Stage Method." Unpublished Ph.D. Thesis, Department of Economics, University of Rochester.
- . 1983. "Generalized Econometric Models with Selectivity." *Econometrica* 51 (2): 507–12.
- Lucas, R E. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets. Carnegie-Rochester Conference Series on Public Policy*, Vol. 1, edited by K. Brunner and A. H. Meltzer, pp. 19–46. Amsterdam: North-Holland.

- Manning, W. G., N. Duan, and W. H. Rogers. 1987. "Monte Carlo Evidence on the Choice between Sample Selection and Two-Part Models." *Journal of Econometrics* 35: 59–82.
- Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel. 2006. "Instrumental Variables: Application and Limitations." *Epidemiology* 17 (3): 260–7.
- McClellan, M., B. McNeil, and J. Newhouse. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association* 272 (11): 859–66.
- Moffitt, R. A. 1996. "Comment." *Journal of the American Statistical Association* 91 (434): 462–5.
- Monfardini, C., and R. Radice. 2008. "Testing Exogeneity in the Bivariate Probit Model: A Monte Carlo study." *Oxford Bulletin of Economics and Statistics* 70 (2): 271–82.
- Murray, M. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments." *Journal of Economic Perspectives* 20 (4): 111–32.
- Olsen, R. 1980. "A Least-Squares Correction for Selectivity Bias." *Econometrica* 48 (7): 1815–20.
- Pearl, J. 1997. "The New Challenge: From a Century of Statistics to an Age of Causation." Computer Science Department, University of California, Los Angeles, CA [accessed on October 22, 2010]. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.2380&rep=rep1&type=>
- . 2000. *Causality: Models, Reasoning and Inference*. Cambridge, U.K.: Cambridge University Press.
- Pearson, K. 1920. "Notes on the History of Correlation." *Biometrika* 13 (1): 25–45.
- Penrose, R. 1989, 2002. *The Emperor's New Mind*. New York: Oxford University Press (citations are from the 1989 edition).
- . 2005, 2007. *The Road to Reality*. New York: Knopf Publishing Group (citations are from the 2005 edition).
- Polanyi, M. 1962. "The Unaccountable Element in Science." *Philosophy: The Journal of the Royal Institute of Philosophy* 37: 1–14.
- . 1968. "Logic and Psychology." *American Psychologist* 23: 27–43.
- Rassen, J. A., M. A. Brookhart, R. J. Glynn, M. A. Mittleman, and S. Schneeweiss. 2009a. "Instrumental Variables I: Instrumental Variables Exploit Natural Variation in Non-Experimental Data to Estimate Causal Relationships." *Journal of Clinical Epidemiology* 62 (12): 1226–1232.
- . 2009b. "Instrumental Variables II: Instrumental Variable Application—In 25 Variations, the Physician Prescribing Preference Generally Was Strong and Reduced Covariate Imbalance." *Journal of Clinical Epidemiology* 62 (12): 1233–1241.
- Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- . 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24.
- Rosenberg, A. 1989. "Russell versus Steiner on Physics and Causality." *Philosophy of Science* 56 (2): 341–7.

- Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.
- Russell, B. 1913. "On the Notion of Cause." *Proceedings of the Aristotelian Society* 13: 1–26.
- Saris, W., and H. Stronkhorst. 1984. *Causal Modeling in Non-Experimental Research*. Amsterdam: Sociometric Research Foundation.
- Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
- Steiner, M. 1986. "Events and Causality." *Journal of Philosophy* 83 (5): 249–64.
- Stock, J. H., and F. Trebbi. 2003. "Retrospectives: Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17 (3): 177–94.
- Stukel, T. A., E. S. Fisher, D. E. Wennberg, D. A. Alter, D. J. Gottlieb, and M. J. Vermeulen. 2007. "Analysis of Observational Studies in the Presence of Treatment Selection Bias: Effects of Invasive Cardiac Management on AMI Survival Using Propensity Score and Instrumental Variable Methods." *Journal of the American Medical Association* 297 (3): 278–85.
- Suppes, P. C. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North Holland.
- Urbach, P. 1985. "Randomization and the Design of Experiments." *Philosophy of Science* 52: 256–73.
- Wright, P. 1928. "The Tariff on Animal and Vegetable Oils: Appendix B." Reproduced in Stock and Trebbi (2003).
- Wright, S. 1918. "On the Nature of Size Factors." *Genetics* 3: 367–74.
- Worrall, J. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69: S216–330.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Two Derivations of the IV Estimator.

Appendix SA2: RCTs and Heterogeneous Treatment Effects.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.