

Ch. 1 in: Gelman, A. and Meng, X.L. (eds.). *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*. New York: Wiley, 2004, pp. 3-13.

## **An Overview of Methods for Causal Inference from Observational Studies**

Sander Greenland

Departments of Epidemiology and Statistics

University of California, Los Angeles

Los Angeles, CA 90095-1772, U.S.A.

[lesdomes@ucla.edu](mailto:lesdomes@ucla.edu)

**Index terms:** bias modeling, causal inference, causal models, cause, confounding, counterfactual, effect, g-estimation, Hill considerations, observational studies, potential outcome, marginal structural model, measurement error, methodologic modeling, missing data, selection bias, structural equation model, structural nested failure-time model.

### **Introduction**

This chapter provides a brief overview of causal-inference methods found in the health sciences. It is convenient to divide these methods into a few broad classes: Those based on formal models of causation, especially potential outcomes; those based on canonical considerations, in which causality is a property of an association to be diagnosed by symptoms and signs; and those based on methodologic modeling. These are by no means mutually exclusive approaches; for example, one may (though need not) base a methodologic model on potential outcomes, and a canonical approach may use modeling methods to address specific considerations. Rather, the categories reflect historical traditions that until recently had only limited intersection.

### **Approaches Based on Causal Models**

#### *Background: Potential Outcomes*

Most statistical methods, from orthodox Neyman-Pearsonian testing to radical subjective Bayesianism, have been labeled by their proponents as solutions to problems of inductive inference (Greenland, 1998), and causal inference may be classified as a prominent (if

not the major) problem of induction. It would then seem that causal-inference methods ought to figure prominently in statistical theory and training. That this has not been so has been remarked on by other reviewers (Pearl, 2000). In fact, despite the long history of statistics to that point, it was not until the 1920s that a formal statistical model for causal inference was proposed (Neyman, 1923), the first example of a *potential-outcome* model.

Skeptical that induction in general and causal inference in particular could be given a sound logical basis, David Hume nonetheless captured the foundation of potential-outcome models when he wrote

“We may define a cause to be an object, followed by another,... where, if the first object had not been, the second had never existed.” (Hume, 1748, p. 115)

A key aspect of this view of causation is its *counterfactual* element: It refers to how a certain outcome event (the “second object,” or effect) would not have occurred if, *contrary to fact*, an earlier event (the “first object,” or cause) had not occurred. In this regard it is no different from standard frequentist statistics (which refer to sample realizations that might have occurred, but did not) and some forms of competing-risk models (those involving a latent outcome that would have occurred, but for the competing risk). This counterfactual view of causation was adopted by numerous philosophers and scientists after Hume (e.g., Mill, 1843; Fisher, 1918; Cox, 1958; Simon and Rescher, 1966; MacMahon and Pugh, 1967; Lewis, 1973).

The development of this view into a statistical theory for causal inference is recounted by Rubin (1990), Greenland et al. (1999), Greenland (2000), and Pearl (2000). To describe that theory, suppose we wish to study the effect of an intervention variable  $X$  with potential values (range)  $x_1, \dots, x_J$  on a subsequent outcome variable  $Y$  defined on an observational unit or a population. The theory then supposes that there is a vector of *potential outcomes*  $\mathbf{y} = (y(x_1), \dots, y(x_J))'$ , such that if  $X = x_j$  then  $Y = y(x_j)$ ; this vector is simply a mapping from the  $X$  range to the  $Y$  range for the unit. To say that intervention  $x_i$  causally affects  $Y$  relative to intervention  $x_j$  then means that  $y(x_i) \neq y(x_j)$ ; and the effect of intervention  $x_i$  relative to  $x_j$  on  $Y$  is measured by  $y(x_i) - y(x_j)$  or (if  $Y$  is strictly positive) by  $y(x_i)/y(x_j)$ . Under this theory, assignment of a unit to a treatment level  $x_i$  is simply a choice of which coordinate of  $\mathbf{y}$  to attempt to observe; regardless of assignment, the remaining coordinates are treated as existing pre-treatment covariates on which data are

missing (Rubin, 1978). Formally, if we define the vector of potential treatments  $\mathbf{x} = (x_1, \dots, x_J)'$ , with treatment indicators  $r_i = 1$  if the unit is given treatment  $x_i$ , 0 otherwise, and  $\mathbf{r} = (r_1, \dots, r_J)'$ , then the actual treatment given is  $x_a = \mathbf{r}'\mathbf{x}$  and the actual outcome is  $y_a \equiv y(x_a) = \mathbf{r}'\mathbf{y}$ . Viewing  $\mathbf{r}$  as the item-response vector for the items in  $\mathbf{y}$ , causal inference under potential outcomes can be seen as a special case of inference under item nonresponse in which  $\sum_i r_i = 0$  or 1, i.e., at most one item in  $\mathbf{y}$  is observed per unit (Rubin, 1991).

The theory extends to stochastic outcomes by replacing the  $y(x_i)$  by probability mass functions  $p_i(y)$  (Greenland, 1987; Robins, 1988; Greenland et al., 1999), so the mapping is from  $X$  to the space of probability measures on  $Y$ . This extension is embodied in the “set” or “do” calculus for causal actions (Pearl, 1995, 2000) described briefly below. The theory also extends to continuous  $X$  by allowing the potential-outcome vector to be infinite-dimensional with coordinates indexed by  $X$ , and components  $y(x)$  or  $p_x(y)$ . Finally, the theory extends to complex longitudinal data structures by allowing the treatments to be different event histories or processes (Robins, 1987, 1997).

#### *Limitations of Potential-Outcome Models*

The power and controversy of this formalization derives in part from defining cause and effect in simple terms of interventions and potential outcomes, rather than leaving them informal or obscure. Judged on the basis of number and breadth of applications, the potential-outcome approach is an unqualified success, as contributions to the present volume attest. Nonetheless, because only one of the treatments  $x_i$  can be administered to a unit, for each unit at most one potential outcome  $y(x_i)$  will become an observable quantity; the rest will remain counterfactual, and hence in some views less than scientific (Dawid, 2000). More specifically, the approach has been criticized for including structural elements that are in principle unidentifiable by randomized experiments alone. An example is the correlation among potential outcomes: Because no two potential outcomes  $y(x_i)$  and  $y(x_j)$  from distinct interventions  $x_i \neq x_j$  can be observed on one unit, nothing about the correlation of  $y(x_i)$  and  $y(x_j)$  across units can be inferred from observing interventions and outcomes alone; the correlation becomes unobservable and hence by some usage “metaphysical.”

This sort of problem has been presented as if a fatal flaw of potential outcomes models (Dawid, 2000). Most commentators, however, regard such problems as indicating inherent limits of inference based on unrepeatable “black-box” observation: For some questions one must go beyond observations of unit responses, to unit-specific investigation of the mechanisms of action (e.g., dissection and physiology). This need is familiar in industrial statistics in the context of destructive testing, although there controversy does not arise because the mechanisms of action are usually well understood. The potential-outcomes approach simply highlights the limits of what statistical analyses can show absent background theory about causal mechanisms, even if treatment is randomized: Standard statistical analyses address only the magnitude of associations and the average causal effects those represent, not the mechanisms underlying those effects.

*Translating Potential Outcomes into Statistical Methodology*

Among the earliest applications of potential outcomes was in randomization tests for causal effects. These applications illustrate the transparency potential outcomes can bring to standard methods, and show their utility in revealing the assumptions needed to give causal interpretations to standard statistical procedures.

Suppose we have  $N$  units indexed by  $n$  and we wish to test the strong (sharp) null hypothesis that treatment  $X$  has no effect on  $Y$  for any unit, i.e., for all  $i, j, n$ ,  $y_n(x_i) = y_n(x_j)$ . Under this null, the observed distribution of  $Y$  among the  $N$  units would not differ from its observed value, regardless of how treatment is allocated among the units. Consequently, given the treatment-allocation probabilities (propensity scores) we may compute the exact null distribution of any measure of differences among treatment groups. In doing so we can and should keep the marginal distribution of  $Y$  at its observed value, for with no treatment effect on  $Y$ , changes in treatment allocation cannot alter the marginal distribution of  $Y$ .

The classic examples of this reasoning are permutation tests based on uniform allocation probabilities across units (simple randomization), such as Fisher’s exact test (Cox and Hinkley, 1974, sec. 6.4). For these tests the fixed  $Y$ -margin is often viewed as a mysterious assumption by students, but can be easily deduced from the potential-outcome formulation, with no need to appeal to obscure and controversial conditionality principles (Greenland, 1991). Potential-outcome models can also be used to derive classical

confidence intervals (which involve non-null hypotheses and varying margins), superpopulation inferences (in which the  $N$  units are viewed as a random sample from the actual population of interest), and posterior distributions for causal effects of a randomized treatment (Robins, 1988; Rubin, 1978). The models further reveal hidden assumptions and limitations of common procedures for instrumental-variable estimation (Angrist et al., 1996), for intent-to-treat analyses (Goetghebeur and van Houwelingen, 1998), for separating direct and indirect effects (Robins and Greenland, 1992, 1994; Frangakis and Rubin, 2002), for confounding identification (Greenland et al., 1999), for estimating causation probabilities (Greenland and Robins, 2000), for handling time-varying covariates (Robins, 1987, 1998b; Robins et al., 1992), and for handling time-varying outcomes (Robins et al., 1999a).

#### *A Case Study: G-Estimation*

Potential-outcome models have contributed much more than conceptual clarification. As documented elsewhere in this volume, they have been used extensively by Rubin, his students, and his collaborators to develop novel statistical procedures for estimating causal effects. Indeed, one defense of the approach is that it stimulates insights which lead not only to the recognition of shortcomings of previous methods, but also to development of new and more generally valid methods (Wasserman, 2000).

Methods for modeling effects of time-varying treatment regimes (generalized treatments, or “g-treatments”) provide a case study in which the potential-outcome approach led to a very novel way of attacking an exceptionally difficult problem. The difficulty arises because a time-varying regime may not only be influenced by antecedent causes of the outcome (which leads to familiar issues of confounding), but may also influence later causes, which in turn may influence the regime. Robins (1987) identified a recursive “g-computation” formula as central to modeling treatment effects under these feedback conditions and derived nonparametric tests based on this formula (a special case of which was first described by Morrison, 1985). These tests proved impractical beyond simple null-testing contexts, which led to development of semiparametric modeling procedures for inferences about time-varying treatment effects (Robins, 1998).

The earliest of these procedures were based on the structural-nested failure-time model (SNFTM) for survival time  $Y$  (Robins et al., 1992; Robins and Greenland, 1994;

Robins, 1998), a generalization of the strong accelerated-life model (Cox and Oakes, 1984). Suppressing the unit subscript  $n$ , suppose a unit is actually given fixed treatment  $X = x_a$  and fails at time  $Y_a = y(x_a)$ , the potential outcome of the unit under  $X = x_a$ . The basic causal accelerated-life model assumes the survival time of the unit when given  $X = 0$  instead would have been  $Y_0 = e^{x_a\beta} Y_a$ , where  $Y_0$  is the potential outcome of the unit under  $X = 0$ , and the factor  $e^{x_a\beta}$  is the amount by which setting  $X = 0$  would have expanded (if  $x_a\beta > 0$ ) or contracted (if  $x_a\beta < 0$ ) survival time relative to setting  $X = x_a$ .

Suppose now  $X$  could vary and the actual survival interval  $S = (0, Y_a)$  is partitioned into  $K$  successive intervals of length  $\Delta t_1, \dots, \Delta t_K$ , such that  $X = x_k$  in interval  $k$ , with a vector of covariates  $Z = z_k$  in the interval. A basic SNFTM for the survival time of the unit had  $X$  been held at zero over time is then  $Y_0 = \sum_k \exp(x_k\beta) \Delta t_k$ ; the extension to a continuous treatment history  $x(t)$  is  $Y_0 = \int_0^{Y_a} \exp(x(t)\beta) dt$ . The model is semiparametric insofar as the distribution of  $Y_0$  across units is not completely specified, although this distribution may be further modeled as a function of baseline covariates.

Likelihood-based inference on  $\beta$  is unwieldy, but testing and estimation can be easily done with a clever two-step procedure called g-estimation (Robins et al., 1992; Robins and Greenland, 1994; Robins, 1998). To illustrate the basic idea, assume no censoring of  $Y$ , no measurement error, and let  $X_k$  and  $Z_k$  be the treatment and covariate random variables for interval  $k$ . Then, under the model, a hypothesized value  $\beta_h$  for  $\beta$  produces for each unit a computable value  $Y_0(\beta_h) = \sum_k \exp(x_k\beta_h) \Delta t_k$  for  $Y_0$ . Next, suppose that for all  $k$   $Y_0$  and  $X_k$  are independent given past treatment history  $X_1, \dots, X_{k-1}$  and covariate history  $Z_1, \dots, Z_k$  (as would obtain if treatment were sequentially randomized given these histories). If  $\beta = \beta_h$ , then  $Y_0(\beta_h) = Y_0$  and so must be independent of  $X_k$  given the histories. One may test this conditional independence of  $Y_0(\beta_h)$  and the  $X_k$  with any standard method. For example, one could use a permutation test or some approximation to one (such as the usual logrank test) stratified on histories; subject to further modeling assumptions, one could instead use a test that the coefficient of  $Y_0(\beta_h)$  is zero in a model for the regression of  $X_k$  on  $Y_0(\beta_h)$  and the histories. In either case,  $\alpha$ -level rejection of conditional independence of  $X_k$  and  $Y_0(\beta_h)$  implies  $\alpha$ -level rejection of  $\beta = \beta_h$ , and the set of all  $\beta_h$  not so rejected form a  $1-\alpha$  confidence set for  $\beta$ . Furthermore, the random

variable corresponding to the value  $b$  for  $\beta$  that makes  $Y_0(b)$  and the  $X_k$  conditionally independent is a consistent asymptotically normal estimator of  $\beta$  (Robins, 1998).

Of course, in observational studies g-estimation shares all the usual limitations of standard methods. The assignment mechanism is not known, so inferences are only conditional on an uncertain assumption of “no sequential confounding;” more precisely, that  $Y_0$  and the  $X_k$  are independent given the treatment and covariate histories used for stratification or modeling of  $Y_0$  and the  $X_k$ . If this independence is not assumed then rejection of  $\beta_h$  only entails that either  $\beta \neq \beta_h$  or that  $Y_0$  and the  $X_k$  are dependent given the histories (i.e., there *is* residual confounding). Also, inferences are conditional on the form of the model being correct, which is not likely to be exactly true, even if fit appears good. Nonetheless, as in many standard testing contexts (such as the classical t-test), under broad conditions the asymptotic size of the stratified test of the no-effect hypothesis  $\beta = 0$  will not exceed  $\alpha$  if  $Y_0$  and the  $X_k$  are indeed independent given the histories (i.e., absent residual confounding), even if the chosen SNFTM for  $Y_0$  is incorrect, although the power of the test may be severely impaired by the model misspecification (Robins, 1998). In light of this “null-robustness” property, g-null testing can be viewed as a natural extension of classical null testing to time-varying-treatment comparisons.

If (as usual) censoring is present, g-estimation becomes more complex (Robins, 1998). As a simpler though more restrictive approach to censored longitudinal data with time-varying treatments, one may fit a marginal structural model (MSM) for the potential outcomes using a generalization of Horvitz-Thompson inverse-probability-of-selection weighting (Robins, 1999; Hernan et al., 2001). Unlike standard time-dependent Cox models, both SNFTM and MSM fitting require special attention to the censoring process, but make weaker assumptions about that process. Thus their greater complexity is the price one must pay for the generality of the procedures, for both can yield unconfounded effect estimates in situations in which standard models appear to fit well but yield very biased results (Robins et al., 1992; Robins and Greenland, 1994; Robins et al., 1999a; Hernan et al., 2001).

#### *Other Formal Models of Causation*

Most statistical approaches to causal modeling incorporate elements formally equivalent to potential outcomes (Pearl, 2000). For example, the sufficient-component cause model

found in epidemiology (Rothman and Greenland, 1998, Ch. 2) is a potential-outcome model. In structural-equation models (SEMs), the component equations can be interpreted as models for potential outcomes (Pearl, 1995, 2000), as in the SNFTM example. The identification calculus based on graphical models of causation (causal diagrams) has a direct mapping into the potential-outcomes framework, and yields the g-computation algorithm as a by-product (Pearl, 1995). These and other connections are reviewed by Pearl (2000) and Greenland and Brumback (2002).

It appears that causal models lacking a direct correspondence to potential outcomes have yet to yield generally accepted statistical methodologies for causal inference, at least within the health sciences. This may represent an inevitable state of affairs arising from a counterfactual element at the core of all commonsense or practical views of causation (Lewis, 1973; Pearl, 2000). Consider the problem of predictive causality: We can recast causal inferences about future events as predictions conditional on specific intervention or treatment-choice events. The choice of  $x$  for  $X$  is denoted “set  $X=x$ ” in Pearl (1995) and “do  $X=x$ ” in Pearl (2000); the resulting collection of predictive probabilities  $P\{Y=y \mid \text{set}(X=x_i)\}$  or  $P\{Y=y \mid \text{do}(X=x_i)\}$  is isomorphic to the set of stochastic potential outcomes  $p_i(y)$ . As Hume (1748) and Lewis (1973) noted, for causal inferences about past events we are typically interested in questions of the form “what would have happened if  $X$  had equaled  $x_c$  rather than  $x_a$ ,” where the alternative choice  $x_c$  does not equal the actual choice  $x_a$  and so must be counterfactual; thus, consideration of potential outcomes seems inescapable when confronting historical causal questions, a point conceded by thoughtful critics of counterfactuals (Dawid, 2000).

### **Canonical Inference**

Some approaches to causal inference bypass definitional controversy by not basing their methods on a formal causal model. The oldest of these approaches is traceable to John Stuart Mill in his attempt to lay out a system of inductive logic based on canons or rules which causal associations were presumed to obey (Mill, 1843). Perhaps the most widely cited of such lists today are the Austin Bradford Hill considerations (misnamed “criteria” by later writers) (Hill, 1965), which are discussed critically in numerous sources (e.g., Koepsell and Weiss, 2003; Phillips and Goodman, 2003; Rothman and Greenland, 1998, Ch. 2), and which will be the focus here.



The canonical approach usually leaves terms like “cause” and “effect” as primitives (formally undefined concepts) around which the self-evident canons are built, much like axioms are built around the primitives of “set” and “is an element of” in mathematics, although the terms may be defined in terms of potential outcomes. Only the canon of proper temporal sequence (cause must precede effect) is a necessary condition for causation. The remaining canons or considerations are not necessary conditions; instead, they are like diagnostic symptoms or signs of causation – that is, properties an association is assumed more likely to exhibit if it is causal than if it is not (Hill, 1965; Susser, 1988, 1991). Thus, the canonical approach makes causal inference appear more akin to clinical judgment than experimental science, although experimental evidence is among the considerations (Hill, 1965; Rothman and Greenland, 1998, Ch. 2; Susser, 1991). Some of the considerations (such as temporal sequence, association, dose-response or predicted gradient, and specificity) are empirical signs and thus subject to conventional statistical analysis; others (such as plausibility) refer to prior belief, and thus (as with disease symptoms) require elicitation, and could be used to construct priors for Bayesian analysis.

The canonical approach is widely accepted in the health sciences, subject to many variations in detail. Nonetheless, it has been criticized for its incompleteness and informality, and the consequent poor fit it affords to the deductive or mathematical approaches familiar to classic science and statistics (Rothman and Greenland, 1998, Ch.2). Although there have been some interesting attempts to reinforce or reinterpret certain canons as empirical predictions of causal hypotheses (e.g., Susser, 1988; Weed, 1986; Weiss, 1981, 2002; Rosenbaum, 2002), there is no generally accepted mapping of the entire canonical approach into a coherent statistical methodology; one simply uses standard statistical techniques to test whether empirical canons are violated. For example, if the causal hypothesis linking X to Y predicts a strictly increasing trend in Y with X, a test of this statistical prediction may serve as a statistical criterion for determining whether the hypothesis fails the dose-response canon. Such usage falls squarely in the falsificationist/frequentist tradition of 20<sup>th</sup> century statistics, but leaves unanswered most of the policy questions that drive causal research (Phillips and Goodman, 2003).

### **Methodologic Modeling**

In the second half of the 20<sup>th</sup> Century a third approach emerged from battles over the policy implications of observational data, such as those concerning the epidemiology of cigarette smoking and lung cancer. One begins with the idea that, conditional on some set of concomitants or covariates  $Z$ , there is a population association or relation between  $X$  and  $Y$  that is the target of inference, usually because it is presumed to accurately reflect the effect of  $X$  on  $Y$  in that population (as in the canonical approach, “cause” and “effect” may be left undefined or defined in other terms such as potential outcomes).

Observational and analytic shortcomings then distort or bias estimates of this effect: Units may be selected for observation in a nonrandom fashion; conditioning on additional unmeasured covariates  $U$  may be essential for the  $X$ - $Y$  association to approximate a causal effect; inappropriate covariates may be entered into the analysis; components of  $X$  or  $Y$  or  $Z$  may not be adequately measured; and so on.

One can parametrically model these methodologic shortcomings and derive effect estimates based on the models. If (as is usual) the data under analysis cannot provide estimates of the methodologic parameters, one can fix the parameters at specific values, estimate effects based those values, and see how effect estimates change as those values are varied (sensitivity analysis). One can also assign the parameters prior distributions based on background information, and summarize the effect estimates over these distributions (e.g., with the resulting posterior distribution). These ideas are well established in engineering and policy research and are covered in many books, albeit in a wide variety of forms and specialized applications. Little and Rubin (2002) focus on missing-data problems; Eddy et al. (1992) focus on medical and health-risk assessment; and Vose (2000) covers general risk assessment. Nonetheless, general methodologic or bias modeling has only recently begun to appear in epidemiologic research (Robins et al., 1999b; Graham, 2000; Gustafson, 2003; Lash and Fink, 2003; Phillips, 2003; Greenland, 2003), although more basic sensitivity analyses have been employed sporadically since the 1950s (see Rothman and Greenland, 1998, Ch. 19, for citations and an overview).

Consider again the problem of estimating the effect of  $X$  on  $Y$ , given a vector of antecedent covariates  $Z$ . Standard approaches are based on estimating  $E(Y|x,z)$  and taking the fitted (partial) regression of  $Y$  on  $X$  given  $Z$  as the effect of  $X$  on  $Y$ . Usually a parametric model  $r(x,z;\beta)$  for  $E(Y|x,z)$  is fit and the coefficient for  $X$  is taken as the effect

(this approach is reflected in common terminology that refers to such coefficients as “main effects”). The fitting is almost always done as if (1) within levels of  $X$  and  $Z$ , the data are a simple random sample and any missingness is completely at random, (2) the causal effect of  $X$  on  $Y$  is accurately reflected by the association of  $X$  and  $Y$  given  $Z$  (i.e., there is no residual confounding – as might be reasonable to assume if  $X$  were randomized within levels of  $Z$ ), and (3)  $X$ ,  $Y$ , and  $Z$  are measured without error. But, in reality, (1) sampling and missing-data probabilities may jointly depend on  $X$ ,  $Y$ , and  $Z$  in an unknown fashion, (2) conditioning on certain unmeasured (and possibly unknown) covariates  $U$  might be essential for the association of  $X$  and  $Y$  to correspond to a causal effect of  $X$  on  $Y$ , and (3)  $X$ ,  $Y$  and  $Z$  components may be mismeasured.

Let  $V = (X, Y, Z)$ . One approach to sampling (selection) biases is to posit a model  $s(v; \sigma)$  for the probability of selection given  $v$ , then use this model in the analysis along with  $r(x, z; \beta)$ , e.g., by incorporating  $s(v; \sigma)$  into the likelihood function (Eddy et al., 1992; Little and Rubin, 2002; Gelman et al., 2003) or by using  $s(v; \sigma)^{-1}$  as a weighting factor (Robins et al., 1994, 1999b). The joint parameter  $(\beta, \sigma)$  is usually not fully identified from the data under analysis, so one must either posit various fixed values for  $\sigma$  and estimate  $\beta$  for each chosen  $\sigma$  (sensitivity analysis), or else give  $(\beta, \sigma)$  a prior density and conduct a Bayesian analysis. A third approach, Monte-Carlo risk analysis or Monte-Carlo sensitivity analysis (MCSA), repeatedly samples  $\sigma$  from its marginal prior, resamples (bootstraps) the data, and re-estimates  $\beta$  using the sampled  $\sigma$  and data; it then gives the distribution of results obtained from this repeated sampling-estimation cycle. MCSA can closely approximate Bayesian results under certain (though not all) conditions (Greenland, 2001, 2004), most notably that  $\beta$  and  $\sigma$  are *a priori* independent and the prior for  $\beta$  is vague. The basic selection-modeling methods can be generalized (with many technical considerations) to handle missing data (Little and Rubin, 2002; Robins et al., 1994, 1999b).

One approach to problem (2) is to model the joint distribution of  $U, V$  with a parametric model  $p(u, v; \beta, \gamma) = p(y|u, x, z; \beta)p(u, x, z; \gamma)$ . Again, one can estimate  $\beta$  by likelihood-based or by weighting methods, but because  $U$  is unmeasured (latent), the parameter  $(\beta, \gamma)$  will not be fully identified from the data and so some sort of sensitivity analysis or prior distribution will be needed (e.g., Yanagawa, 1984; Robins et al., 1999b;

Greenland, 2003, 2004). Results will depend heavily on the prior specification given  $U$ . For example,  $U$  may be a specific unmeasured covariate (e.g., smoking status) with well studied relations to  $X$ ,  $Y$ , and  $Z$ , which affords straightforward Bayesian and MCSA analyses (Steenland and Greenland, 2004). On the other hand,  $U$  may represent an unspecified aggregation of latent confounders, in which case the priors and hence inferences are more uncertain (Greenland, 2003).

Next, suppose that the “true” variable vector  $V = (X, Y, Z)$  has the corresponding measurement or surrogate  $W$  (a vector with subvectors corresponding to measurements of components of  $X$ ,  $Y$ , and  $Z$ ). The measurement-error problem (problem 3) can then be expressed as follows: For some or all units, at least one of the  $V$  components is missing, but the measurement (subvector of  $W$ ) corresponding to that missing  $V$  component is present. If enough units are observed with both  $V$  and  $W$  complete, the problem can be handled by standard missing-data methods. For example, given a model for the distribution of  $(V, W)$  one can use likelihood-based methods (Little and Rubin, 2002), or impute  $V$  components where absent and then fit the model  $r(x, z; \beta)$  for  $E(Y|x, z)$  to the completed data (Cole et al., 2004), or fit the model to the complete records using weights derived from all records using a model for missing-data patterns (Robins et al., 1994, 1999b). Alternatively, there are many measurement-error correction procedures that directly modify  $\beta$  estimates obtained by fitting the regression using  $W$  as if it were  $V$ ; this is usually accomplished with a model relating  $V$  to  $W$  fitted to the complete records (Ruppert et al., 1995).

If a component of  $V$  is never observed on any unit (or, more practically, if there are too few complete records to support large-sample missing-data or measurement-error procedures), one may turn to latent-variable methods (Berkane, 1997). For example, one could model the distribution of  $(V, W)$  or a sufficient factor from that distribution by a parametric model; the unobserved components of  $V$  are the latent variables in the model. The parameters will not be fully identified, however, and sensitivity analysis or prior distributions will again be needed. In practice a realistic specification can become quite complex, with subsequent inferences displaying extreme sensitivity to parameter constraints or prior distribution choices (e.g., Greenland, 2004). Nonetheless, display of

this sensitivity can help provide an honest accounting for the large uncertainty that can be generated by apparently modest and realistic error distributions.

### Conclusion

The three approaches described above represent separate historical streams rather than distinct methodologies, and can be blended in various ways. For example, methodologic models for confounding or randomization failure are often based on potential outcomes; the result of any modeling exercise is simply one more input to larger, informal judgments about causal relations; and those judgments may be guided by canonical considerations. Insights and innovations in any approach can thus benefit the entire process of causal inference, especially when that process is seen as part of a larger context. Finally, other traditions or approaches (some perhaps yet to be imagined) may contribute to the process. Hence I would advise against regarding any one approach or blending as a complete solution or algorithm for problems of causal inference; the area remains one rich with open problems and opportunities for innovation.

### References

- Angrist, J., Imbens, G. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, **91**, 444-472.
- Berkane, M., ed. (1997). *Latent Variable Modeling and Applications to Causality*. Lecture Notes in Statistics (120), New York: Springer Verlag.
- Cole, S.R., Chu, H. and Greenland, S. (2004). Using multiple-imputation for measurement error correction in pediatric chronic kidney disease. *American Journal of Epidemiology*, **159**, to appear.
- Cox, D.R. (1958). *The Planning of Experiments*. New York: Wiley.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. New York: Chapman and Hall.
- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.
- Dawid, P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, **95**, 407-448.
- Eddy, D.M., Hasselblad, V. and Schachter, R. (1992). *Meta-Analysis by the Confidence Profile Method*. New York: Academic Press.
- Fisher, R.A. (1918). The causes of human variability. *Eugenics Review*, **10**, 213-220.
- Frangakis, C. and Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21-29.
- Goetghebeur, E. and van Houwelingen, H.C., eds. (1998). Analyzing noncompliance in clinical trials. *Statistics in Medicine*, **17**, 247-389.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*, 2<sup>nd</sup> ed. New York: Chapman and Hall/CRC.
- Graham, P. (2000). Bayesian inference for a generalized population attributable fraction.

- Statistics in Medicine* **19**, 937-956.
- Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analysis. *American Journal of Epidemiology*, **125**, 761-768.
- Greenland, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *The American Statistician*, **45**, 248-251.
- Greenland, S. (1998). Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology*, **27**, 543-548.
- Greenland, S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association*, **95**, 286-289.
- Greenland, S. (2001). Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis*, **21**, 579-583.
- Greenland, S. (2003). The impact of prior distributions for uncontrolled confounding and response bias. *Journal of the American Statistical Association*, **98**, 47-54.
- Greenland S. (2005). Multiple-bias modeling for observational studies (with discussion). *Journal of the Royal Statistical Society*, series A, to appear.
- Greenland, S., Robins, J. M. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, **14**, 29-46.
- Greenland, S. and Robins, J.M. (2000). Epidemiology, justice, and the probability of causation. *Jurimetrics*, **40**, 321-340.
- Greenland, S. and Brumback, B.A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, **31**, 1030-1037.
- Gustafson P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology*. New York: Chapman and Hall.
- Hernan, M.A., Brumback, B.A. and Robins, J.M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, **96**, 440-448.
- Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295-300.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. 1988 reprint by Open Court Press.
- Lash, T.L. and Fink, A. K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. *Epidemiology*, **14**, 451-458.
- Lewis, D.K. (1973). Causation. *Journal of Philosophy*, **70**, 556-567.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2<sup>nd</sup> ed. New York: Wiley.
- MacMahon, B. and Pugh, T.F. (1967). Causes and entities of disease. In Clark, D.W. and MacMahon, B. *Preventive Medicine*. Boston: Little, Brown, 11-18.
- Mill, J.S. (1843). *A System of Logic, Ratiocinative and Inductive*. 1956 reprint by Longman and Greens, London.
- Morrison, A.S. (1985). *Screening in Chronic Disease*. New York: Oxford.
- Neyman, J. (1923). Sur les applications de la thar des probabilités aux expériences Agaricales: Essay des principe [English translation of excerpts (1990) by D. Dabrowska and T. Speed, *Statistical Science*, **5**, 463-472.]
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669-710.
- Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
- Phillips, C V. (2003). Quantifying and reporting uncertainty from systematic errors. *Epidemiology*, **14**, 459-466.
- Phillips, C.V. and Goodman, K. (2003). The messed lessons of Sir Austin Bradford Hill.

- [www.epi.phh.com/papers/phillips-goodman\\_abhill\\_nov03.pdf](http://www.epi.phh.com/papers/phillips-goodman_abhill_nov03.pdf).
- Robins, J.M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, **40**, supplement 2, 139s-161s.
- Robins, J.M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, **7**, 773-785.
- Robins, J.M. (1997). Causal inference from complex longitudinal data. In: Berkane, M., ed. *Latent Variable Modeling and Applications to Causality*, 69-117. Lecture Notes in Statistics (120), New York: Springer Verlag.
- Robins, J.M. (1998). Structural nested failure time models. In: Armitage, P. and Colton, T. (eds.). *The Encyclopedia of Biostatistics*. New York: Wiley, 4372-4389.
- Robins J.M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, M.E. and Berry D.A., eds. *Statistical Models in Epidemiology*. New York: Springer-Verlag, 95-134.
- Robins, J.M., Blevins, D., Ritter, G. and Wulfson, M. (1992). G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology*, **3**, 319-336. Errata: *Epidemiology*, **4**, 189.
- Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143-155.
- Robins, J.M. and Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, **89**, 737-749.
- Robins, J.M., Greenland, S. and Hu, F.C. (1999a). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**, 687-712.
- Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.
- Robins, J.M., Rotnitzky, A. and Scharfstein, D.O. (1999b). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, M.E. and Berry D.A., eds. *Statistical Models in Epidemiology*. New York: Springer-Verlag, 1-92.
- Rosenbaum, P. (2002). *Observational Studies*. New York: Springer.
- Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology*, 2<sup>nd</sup> ed. Philadelphia: Lippincott.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701.
- Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, **6**:34-58.
- Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**, 472-480.
- Rubin, D.B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, **47**, 1213-1234.
- Ruppert, D., Stefanski, L.A. and Carroll, R.J. (1995). *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.
- Simon, H.A. and Rescher, N. (1966). Cause and counterfactual. *Philosophy of Science*, **33**, 323-340.
- Steenland, K. and Greenland, S. (2004). Monte-Carlo sensitivity analysis and Bayesian

- analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*, **160**, in press.
- Susser, M. (1988). Falsification, verification and causal inference in epidemiology: reconsideration in light of Sir Karl Popper's philosophy. In: Rothman, K.J. ed. *Causal Inference*. Boston: Epidemiology Resources, Inc., 33-57.
- Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology*, **133**, 635-648.
- Vose, D. (2000). *Risk Analysis*. New York: John Wiley and Sons.
- Wasserman, L. (2000). Comment. *Journal of the American Statistical Association*, **95**, 442-443.
- Weed, D.L. (1986). On the logic of causal inference. *American Journal of Epidemiology*, **123**, 965-979.
- Weiss, N.S. (1981). Inferring causal relationships: elaboration of the criterion of "dose-response." *American Journal of Epidemiology*, **113**, 487-490.
- Weiss, N.S. (2002). Can "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, **13**, 6-8.
- Yanagawa, T. (1984). Case-control studies: assessing the effect of a confounding factor. *Biometrika*, **71**, 191-194.