

James M. ROBINS and Sander GREENLAND

---

By narrowly concentrating on randomized experiments with complete compliance, Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on “decision analysis” and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes. We argue that when attempting to estimate the effects of causes in observational studies or in randomized experiments with noncompliance (termed broken experiments by Barnard et al. (1998),

reliance on counterfactuals or their logical equivalents cannot be avoided.

Causal inference from observational data and broken experiments historically has been viewed as problematic, and even illegitimate, by most statisticians. Thus we regard it as a serious oversight for Dawid to deny the usefulness of a counterfactuals without a more careful consideration of observational studies and broken experiments. The purpose of this discussion is to redress that oversight, by reviewing the considerations that have led

---

James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. Sander Greenland is Professor of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095.

---

© 2000 American Statistical Association  
Journal of the American Statistical Association  
June 2000, Vol. 95, No. 450, Theory and Methods

so many to adopt a counterfactual approach to causal inference.

## 1. THE PROBLEM OF CONFOUNDING

Suppose that we have discrete pretreatment covariates  $A$  and  $B$  in an observational study. At each level of  $A$ , suppose that treatment  $T$  taking values in  $\{t, c\}$  is positively correlated with a disease outcome  $Y$ , but at each joint level  $AB$  of  $A$  and  $B$ , treatment is independent of outcome. In the language of the school of probabilistic causality (PC),  $AB$  screens off  $T$  from  $Y$  (Suppes 1970). Some PC texts would then say that treatment does not probabilistically cause  $Y$  relative to the causal field determined by  $A$  and  $B$ . However, this statement does not reflect the common language and appropriate policy meaning of a cause, which is that manipulating  $T$  would change  $Y$ . Indeed, there is a potential for an infinite regress wherein the association of  $T$  and  $Y$  varies among positive, negative, and null as one adjusts for additional covariates.

In epidemiology, it has been common to view the association adjusted for all measured pretreatment covariates as most likely to be causal. But Greenland and Robins (1986) noted that additional adjustment can increase confounding, in that the more adjusted association could be further from the true average causal effect than the less adjusted association. This problem has also been noted in the PC literature. As a result, the most sophisticated PC texts state that an adjusted effect is guaranteed to have a causal interpretation only when one has succeeded in adjusting for all nontreatment causes  $X$  of the outcome. It then follows as a theorem that the association of treatment and the outcome within levels of the measured covariates, say  $W$ , has a causal interpretation if either (a) the other elements  $X \setminus W$  of  $X$  are independent of  $T$  given  $W$  or (b)  $X \setminus W$  is independent of  $Y$  given  $W$  and  $T$  (Robins and Morgenstern 1987).

Unfortunately, these sufficient "conditions for no confounding" are never empirically testable from observational data, because by definition  $X$  contains all nontreatment causes, including those unmeasured and those not even known to exist. Hence the question of the existence and magnitude of confounding by the unmeasured factors  $X \setminus W$  in an observational study is metaphysical in Dawid's sense, even under his preferred PC theory. It follows that causal inference from observational data is a Dawidian goat. In more standard statistical parlance, the average causal effect of a treatment is not identified from observational data without making nonidentifiable assumptions about the magnitude and direction of confounding. We are confident that Dawid does not wish to join R. A. Fisher (1959) in thereby concluding that causal inferences from observational data are illegitimate, including the inference that cigarette smoking is a cause of lung cancer (Stolley 1991). If we are correct, then Dawid has no choice but to recognize the need for untestable assumptions.

In an attempt to stave off the need for untestable assumptions, some commentators have argued that one should consider as potential confounders only those (often few) variables for which one can make a plausible case that they are

common causes of treatment and the outcome. We find this reasoning unacceptable. Not only does it make confounding a property of the mind rather than of the physical world, but it rewards ignorance. The less one knows about possible causes, the freer one is to make definitive causal statements. The price of this freedom is that more, if not most, of these statements will be false.

### 1.1 Counterfactuals

As Dawid recognizes, in a deterministic (i.e., Newtonian or Laplacian) world with a single time-independent treatment  $T$ , the "all causes" approach to causal inference implies the existence of counterfactuals: If the world is deterministic (i.e., fatalistic) and  $X = X(u)$  includes all nontreatment causes for subject  $u$ , then the outcome must be a deterministic function  $f(i, X(u))$  of  $X(u)$  and the treatment  $i \in \{t, c\}$ . We can then define the counterfactual  $Y_i = Y_i(u)$  to be  $f(i, X(u))$ . In the general case with time-varying treatments, covariates, and outcomes, Robins (1995a, 1997) proved that Pearl's "all causes" nonparametric structural equation model is mathematically equivalent to a special case of the general counterfactual causal model of Robins (1986, 1987) (see also Galles and Pearl 1997). Indeed, the counterfactuals  $Y_i(u)$ ,  $i \in \{t, c\}$  are exactly the ultimate covariates needed for adjustment. Because  $Y = Y(u)$  is a deterministic function of  $(T(u), Y_t(u), Y_c(u))$ , all other variables are independent of  $Y(u)$  given treatment  $T = T(u)$  and  $(Y_t(u), Y_c(u))$ .

Allowing stochastic counterfactuals as done by Robins (1986, 1988), Greenland (1987), and Robins and Greenland (1989, 1991), we can show that even in nondeterministic settings, the "all causes" approach implies the existence of counterfactuals. Hence we reject Dawid's argument that the "all causes" approach is less metaphysical than the counterfactual approach because of the latter's reliance on "complementary" variables. To be specific, suppose that  $Y(u)$  is Bernoulli. Consider a stochastic counterfactual model with the following properties: (a) There exist counterfactual probabilities  $(\theta_t, \theta_c) = (\theta_t(u), \theta_c(u))$  that are deterministic functions of the individual  $u$ ; (b) the function  $Y_i(u)$  is the outcome of a Bernoulli experiment with success probability  $\theta_i(u)$  when  $T(u) = i$ ;  $Y_i(u)$  is undefined when  $T(u) \neq i$ ; and (c)  $Y(u) = Y_{T(u)}(u)$ . This model implies that  $\theta_t(u)$  and  $\theta_c(u)$  have a joint distribution but  $Y_t(u)$  and  $Y_c(u)$  do not. If we take the "all causes" approach as a primitive, we can, in complete parallel with our argument in the deterministic case, define the counterfactuals  $\theta_i(u)$  to be the deterministic function  $f(i, X(u))$  for which (a)–(c) hold.

### 1.2 Stochastic Versus Deterministic Worlds

The deterministic counterfactual model is the limiting special case of the stochastic in which  $\theta_t(u)$  and  $\theta_c(u)$  are always either 1 or 0. As it is impossible to use observational data to empirically decide whether the world is deterministic versus stochastic, we now investigate the inferential consequences of this inability.

**1.2.1 No Unmeasured Confounders.** Let  $W$  denote the measured pretreatment covariates. In a counterfactual

model, we say there are no unmeasured confounders if  $\theta_i \perp\!\!\!\perp T|W$  for  $i \in \{t, c\}$ . This assumption will always hold in a randomized experiment with complete compliance. Given no unmeasured confounders, the marginal distributions  $P_c$  and  $P_t$  of  $Y_c$  and  $Y_t$  given  $W$  are identified and equal to the distributions of  $Y$  given  $W$  and  $T = c$  and  $T = t$ . Thus, as discussed by Dawid and by Robins (1986), if the goal is to determine treatment for a subject  $u_0$  exchangeable with the study subjects by comparing  $P_t$  to  $P_c$ , then it does not matter whether the world is stochastic or deterministic.

We agree with Dawid's concern that an analyst may obtain inconsistent estimates of  $P_t$  and  $P_c$  by specifying a parametric model for nonidentifiable features of the joint distribution of  $(Y_t, Y_c)$ . Our conclusion is not to reject counterfactuals models, however, but rather to criticize models and measures of effect that depend on nonidentifiable features (Greenland 1987) and to develop semiparametric counterfactual models (i.e., structural nested models, marginal structural models, and models based on the  $g$ -computation algorithm) that place no restrictions on those features (Robins 1997, 1999). Our approach completely obviates Dawid's concern.

**1.2.2 Unmeasured Confounders.** Because of the potential for confounding by unmeasured factors, causal effects are not identified by observational data, and the distribution of those data only implies bounds on the causal effect. For deterministic counterfactual models, the bounds always include the causal null hypothesis (Robins 1989). For the stochastic model in which for each  $i \in \{t, c\}$ , the  $\theta_i(u)$  have the same value for all subjects  $u$  within a stratum of the measured covariates, there is no possibility of confounding, association is causation, and the upper and lower bounds coincide. Other assumptions concerning the joint distribution of  $(T, \theta_t, \theta_c)$  will result in bounds intermediate in length. Because whether the world is deterministic is not testable, any value lying within the deterministic bounds can never be rejected by the data. When bounds are too wide to be useful, other approaches to incorporating uncertainty due to unmeasured confounding include sensitivity analysis and formal Bayesian inference (Robins, Scharfstein, and Rotnitzky 1999). As with bounds, the resulting inferences may depend on whether one specifies a deterministic versus a stochastic counterfactual model.

**1.2.3 Counterfactual Analyses That Make a Fundamental Use of Determinism.** Dawid notes that in certain counterfactual analyses, the causal contrasts of interest may have no meaning if the world is stochastic. Dawid cites Imbens and Rubin (1997) for one example. The counterfactual analysis of death as a competing risk by Robins (1986, remark 12.2; 1995b) is a second. We describe a simplified single-occasion discrete-time version of Robins's analysis and provide a new approach that yields meaningful causal contrasts in both deterministic and stochastic worlds.

**Example: Competing risks in a deterministic world.** We observe data  $(ZY, Y, T)$ , where  $T = T(u)$  is a randomized treatment,  $Y = Y(u) = 1$  if subject  $u$  is alive at 6 months and  $Y(u) = 0$  otherwise, and  $Z = Z(u)$  is blood

pressure measured at 6 months, which is observed only if  $Y(u) = 1$ . We refer to death as a "competing risk" for the ability to observe  $Z(u)$ . In the literature, the counterfactuals  $(Z_i(u), Y_i(u)), i \in \{t, c\}$ , are often assumed to exist, in which case average causal effect of treatment on blood pressure is  $E[Z_t(u) - Z_c(u)]$ . This assumption implies that blood pressure  $Z_i(u)$  at 6 months under treatment  $i$  is defined (although never observable) even though the subject  $u$  would be dead; that is,  $Y_i(u) = 0$ . Odd though it may seem, this may sometimes be a useful assumption; for example, if we were studying young children in a developing country. It would be much less reasonable if we were studying adults for whom hypertension is an important cause of death. Even when assumed to be well defined, the measure  $E[Z_t(u) - Z_c(u)]$ , like the other measures of the effect of treatment on blood pressure considered later, is not nonparametrically identified from the data  $(ZY, Y, T)$ ; the distribution of the data only imply bounds for the measure. In contrast, an effect measure relevant for choosing the optimal treatment under a particular utility function for the joint outcome  $(ZY, Y)$  will be identifiable. Nonetheless, a basic scientist's interest may lie in the unidentified effect of treatment on blood pressure.

Kalbfleisch and Prentice (1980) argued that it was never sensible to view  $Z_i(u)$  as well-defined function of  $u$  if  $Y_i(u) = 0$ , in which case  $E[Z_i(u) - Z_c(u)]$  is undefined as well. In that case, Robins (1986) noted that a meaningful measure of the effect of treatment on blood pressure would be its effect  $\Delta_{tc} = E[Z_t(u) - Z_c(u)|Y_c(u) = Y_t(u) = 1]$  on subjects who would survive to 6 months under either treatment. This definition has two drawbacks. First, as noted by Robins (1986), it can result in nontransitivity of treatment comparisons when the treatment has three or more levels. For example, if  $T$  has support  $\{t, c, r\}$ , then it is possible that  $\Delta_{tc}, \Delta_{cr}$ , and  $\Delta_{rt}$  are all positive, so that  $t$  is "preferred" to  $c$ ,  $c$  is preferred to  $r$ , and  $r$  is preferred to  $t$ . Transitivity can be restored by replacing the measure  $\Delta_{tc}$  by  $\Delta_{tc}^* = E[Z_t(u) - Z_c(u)|\{Y_i(u) = 1; i \in \text{support}(T)\}]$ , but then the probability of being in the conditioning set may be small or even 0 if the support of  $T$  is big.

**A Stochastic World Generalization.** The world may be stochastic. Under the stochastic counterfactual model of Section 1.1,  $Y_c(u)$  and  $Y_t(u)$  do not have a joint distribution, but unless they do,  $\Delta_{tc}$  is without meaning. One solution is to add to our stochastic counterfactual model the assumption that  $Y_c$  and  $Y_t$  have a joint distribution. For the model to continue to satisfy properties analogous to (a)–(c), we assume that, with  $(\theta_c, \theta_t)$  as in Section 1.1, the conditional density  $f(Y_c, Y_t|\theta_c, \theta_t)$  factors as  $f(Y_c|\theta_c)f(Y_t|\theta_t)$ , so that  $Y_c$  and  $Y_t$  are independent given  $(\theta_c, \theta_t)$ . Further, we impose the restriction that  $Z_i \perp\!\!\!\perp Y_j|Y_i = 1, \theta_c, \theta_t$  for  $i, j \in \{t, c\}$ , reflecting the fact that  $Y_j$  is purely random given  $(\theta_c, \theta_t)$ . This model is similar to that in Dawid's Section 12. Under this model, it is an elementary calculation to show that  $\Delta_{tc} = E^*[\phi_{tc}(u)]$ , where  $\phi_{tc}(u) \equiv \phi_{tc}(\theta_c(u), \theta_t(u))$  is the random variable

$$\begin{aligned} \phi_{tc}(u) \equiv & E[Z_t(u)|\theta_c(u), \theta_t(u), Y_t(u) = 1] \\ & - E[Z_c(u)|\theta_c(u), \theta_t(u), Y_c(u) = 1] \end{aligned}$$

and  $E^*[\cdot]$  denotes an expectation taken with respect to the weighted density  $f^*(\theta_c, \theta_t) \propto \theta_c \theta_t f(\theta_c, \theta_t)$ .

This approach has two deficiencies when the world is truly stochastic. First, it is no longer logically necessary to define the effect of treatment only for the (possibly quite small) subset with  $Y_c(u) = Y_t(u) = 1$ . Second, in assuming a joint distribution for  $Y_c(u)$  and  $Y_t(u)$ , the approach fails to satisfy Dawid's desire to keep metaphysical (i.e., untestable) assumptions to a minimum. The following alternative solution overcomes both deficiencies. We take  $\phi_{tc}(u)$  as the definition of the causal effect of treatment on subject  $u$ 's blood pressure whenever  $\phi_{tc}(u)$  is defined; that is,  $\theta_t(u)\theta_c(u) \neq 0$ . For subjects for whom  $\theta_t(u)\theta_c(u) = 0$ , we leave the causal effect undefined. Then  $\Phi_{tc} = E[\phi_{tc}(u)I\{\theta_t(u)\theta_c(u) \neq 0\}]/\text{pr}\{\theta_t(u)\theta_c(u) \neq 0\}$  is the average causal effect of treatment on blood pressure among all subjects  $u$  for whom the causal effect is defined, where  $I(\cdot)$  is the indicator function. On the one hand, suppose the world is deterministic. Then, as required,  $\Phi_{tc} = \Delta_{tc}$ . On the other hand, suppose the world is "fully stochastic" in the sense that  $\theta_t(u)\theta_c(u) \neq 0$  for all  $u$ , and it makes sense to regard  $Z_i(u)$  as defined even if  $Y_i(u) = 0$ . Then  $\Phi_{tc} = E[Z_t(u) - Z_c(u)]$ , when we assume that  $Z_i \perp\!\!\!\perp Y_i | \theta_c, \theta_t$  for  $i \in \{t, c\}$  so as to reflect the  $Y_i$ 's being purely random given  $(\theta_c, \theta_t)$ . Thus the approach to the problem of competing risks based on our alternative solution yields all previously proposed measures for the effect of treatment on blood pressure as special cases.

## 2. COUNTERFACTUALS, VAGUENESS, AND OBSERVATIONAL STUDIES

Historically, the main criticism of counterfactuals has not been the statistical objection to positing a joint distribution for complementary variables, but rather the incontrovertible fact that most counterfactuals are inherently vague or ill-defined. We argue, however, that, to misquote the Bard, "the vagueness is not in our counterfactuals but in our attempt to make causal inferences from observational data." To forswear vagueness is to join with Fisher and forswear causal inference from nonexperimental data.

The following proposition of Quine's (1950) effectively ended counterfactual analysis among philosophers until the 1960s: If Bizet and Verdi had been of the same nationality, they both would have been French. Quine argued that because Bizet was French and Verdi was Italian, by symmetry considerations, this counterfactual could not have a truth value and thus was an ill-defined proposition. David Lewis (1973) rejoined that even though some counterfactual propositions may be ill-defined and nearly all are somewhat vague, many are useful. We agree. In fact, we believe that counterfactuals are "vague" precisely to the degree to which one fails to make precise the hypothetical interventions and the causal contrasts under consideration. For example, suppose that one collects observational data to examine the hypothesis that drinking alcohol protects against heart disease. Alcohol may protect against heart disease via a variety of pathways: It may have a direct effect on blood lipid composition; it may relax type A personalities, thereby decreasing

stress-induced hypertension; it may stimulate liver enzymes that detoxify cardiac toxins such as cigarette smoke; it may displace in the diet other items, such as rich desserts, that themselves cause heart disease. If the causal contrast of interest is the direct biological effect of alcohol not mediated through its effect on diet, then one might compare an intervention wherein the daily consumption of alcohol is set to 200 kilocalories (about 2 ounces) and the diet is fixed at a prespecified menu to one in which alcohol consumption is prevented and diet is again held to the same menu.

If, however, alcohol delivered in spirits could have a different effect from alcohol delivered in wine, then these interventions must also specify the source of alcohol. Like attempts to specify all potential common causes (confounders), any attempt to eliminate all vagueness from our interventions leads to an infinite regress wherein we need to specify the type of wine, the vineyard, the year, and other factors. On the way, we eliminate the relevance of any empirical data to our causal query. For example, we might have available disaggregated data on wine and spirit consumption, but similar data on vineyard and year are out of the question.

Only in a randomized experiment in which the interest lies in the causal effect of the entire protocol (so that problems of noncompliance and lack of double-blindings are irrelevant) can we succeed in eliminating all vagueness. In observational studies, the source of the vagueness is the fundamental unavoidable difficulty in formulating just what it is we mean by the causal effect of alcohol on heart disease; the vagueness of counterfactuals is a symptom, not the cause of this difficulty. Dawid appears to express closely related sentiments in his Section 14. Thus we were surprised by Dawid's comment in Section 10 that the two appendixes of Greenland et al. (1999) were convincing illustrations of the meaninglessness and pointlessness of counterfactuals, for we can only interpret his comment as saying that causal inference from nonexperimental data is meaningless and pointless.

## 3. TESTABILITY AND POPPER

Contrary to Dawid's comments, the fact that counterfactual models have untestable elements does not make them "unscientific" according to either the philosophy of Popper or more modern philosophies of science. Popper made clear that falsifiability means a theory must have *some* observable predictions that would lead to its rejection were those predictions to fail, not that *every* feature of the theory be testable (Popper 1974). Counterfactual causal theories meet this requirement by having testable (observable) consequences for the marginal outcome distributions in randomized trials with complete compliance. That observational data do not always provide such critical tests is an inherent difficulty with the data source, not with the theory. Popper also made clear that "metaphysical" (i.e., apparently untestable) elements of theories could be scientifically important in providing guidance for the further development of both theory and method (Popper 1982). From this perspective, as Dawid recognizes, the counterfactual approach

has already shown itself to be an invaluable metaphysical research program for causal inference from observational studies. Similarly, counterfactuals play a key role in several speculative interpretations of quantum phenomena (e.g., Penrose 1994, pp. 237–306; Price 1996, pp. 132–194).

In summary, we regard counterfactuals as a powerful tool for eliminating, to the extent possible, vagueness as to the causal contrasts and hypothetical interventions under consideration. They do so by requiring interested parties to explicate the scientifically important features of the “closest possible worlds” in which all subjects receive or do not receive treatment. Although presented in the context of explicating the difficulty of estimating the causes of effects rather than the effects of causes in observational studies and broken experiments, we agree that many of the points made by Dawid in Sections 11–14 are genuine and difficult problems, and we have considered them in this discussion as well as in our other writings. We believe that these problems are fundamental problems of causal inference that can either be revealed or concealed by a causal theory but never eliminated. Because counterfactuals force these problems into the open, we regard Dawid’s essay as a “shoot the messenger” response to counterfactual theory.

#### ADDITIONAL REFERENCES

- Barnard, J., Du, J., Hill, J. L., and Rubin, D. B. (1998), “A Broader Template for Analyzing Broken Randomized Experiments,” *Sociological Methods and Research*, 27, 285–317.
- Fisher, R. A. (1959), *Smoking—The Cancer Controversy: Some Attempts to Assess The Evidence*, Edinburgh: Oliver and Boyd.
- Galles, D., and Pearl, J. (1998), “An Axiomatic Characterization of Causal Counterfactuals,” *Foundations of Science*, 3, 151–182.
- Greenland, S. (1987), “Interpretation and Choice of Effect Measures in Epidemiologic Analysis,” *American Journal of Epidemiology*, 125, 761–768.
- Greenland, S., and Robins, J. M. (1986), “Identifiability, Exchangeability, and Epidemiologic Confounding,” *International Journal of Epidemiology*, 15, 413–419.
- Kalbfleisch, J. D., and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.
- Penrose, R. (1994), *Shadows of the Mind*, New York: Oxford University Press.
- Popper, K. R. (1974), “The Problem of Demarcation,” in *Popper Selections*, ed. D. M. Miller, Princeton, NJ: Princeton University Press, pp. 101–117.
- (1982), “A Metaphysical Epilogue,” in *Quantum Theory and The Schism in Physics*, (ed. K. R. Popper, Totowa, NJ: Rowman and Littlefield, chap. 4.
- Price, H. (1996), *Time’s Arrow and Archimedes’ Point*, New York: Oxford University Press.
- Quine, W. V. (1950), *Methods of Logic*, New York: Holt, Reinhardt, and Winston.
- Robins, J. M. (1988), “Confidence Intervals for Causal Parameters,” *Statistics in Medicine*, 7, 773–785.
- (1989), “The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies,” in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service, National Center for Health Services Research, pp. 113–159.
- (1995a), Discussion of “Causal Diagrams for Empirical Research” by J. Pearl, *Biometrika*, 82, 695–698.
- (1995b), “An Analytic Method for Randomized Trials With Informative Censoring: Part I,” *Lifetime Data Analysis*, 1, 241–254.
- (1997), “Causal Inference From Complex Longitudinal Data,” in *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, New York: Springer-Verlag, pp. 69–117.
- (1999), “Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference,” in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag, pp. 95–134.
- Robins, J. M., and Greenland, S. (1991), “Estimability and Estimation of Years of Life Lost Due to a Hazardous Exposure,” *Statistics in Medicine*, 10, 79–93.
- Robins, J. M., and Morgenstern, H. (1987), “The Foundations of Confounding in Epidemiology,” *Computers and Mathematics With Applications*, 14, 869–916.
- Robins, J. M., Scharfstein, D., and Rotnitzky, A. (1999), “Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models,” in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag, pp. 1–94.
- Stolley, P. D. (1991), “When Genius Errs,” *American Journal of Epidemiology*, 133, 416–425.
- Suppes, P. (1970), *A Probabilistic Theory of Causation*, Amsterdam: North-Holland.