Statistics and Causal Inference: Rejoinder

Paul W. Holland

Journal of the American Statistical Association, Vol. 81, No. 396. (Dec., 1986), pp. 968-970.

able to place every unit under every value of the controlled variable at every moment of time becomes less plausible. We are back to not being able to relate race or sex to income. Because of the memory, it seems that all such experiments become strictly impossible, as what happened in the past will potentially affect the outcome of the present experiment. It seems to me that most of the solutions to what Holland calls the "Fundamental Problem of Causal Inference" will no longer work in this case, including the "statistical solution," without conditioning on the past. I am thus unclear that the experimental model is even theoretically helpful for temporal causality in the behavioral sciences. If one does condition on the past, the statistical solution may be relevant, but the basis for the inference will then be quite different from that proposed here.

ADDITIONAL REFERENCE

Granger, C. W. J. (in press), "Causality Testing in a Decision Science," to appear in proceedings of the "Conference on Probability and Causality," held at the University of California, Irvine, July 1985.

Rejoinder

PAUL W. HOLLAND

I thank all of the discussants for their very thoughtful comments. Not surprisingly, I agree more with the views expressed by Cox and Rubin than with those of Glymour and Granger, but each discussant makes important points that expand and illuminate issues that arise in the article. Space does not permit a response to every point mentioned, and the more critical comments of Glymour and Granger tend to be balanced by the comments of Cox and Rubin. Hence I will restrict my rejoinder to those issues that I feel need emphasis or to which I feel I can add a useful point of view.

In reflecting upon the discussants' remarks I realized that nowhere in the article, or elsewhere, is there a purely mathematical description of Rubin's model. Such a formulation ought to help separate the *model* itself from its *applications*. For this reason I will begin my rejoinder with a brief, mathematical statement of Rubin's model and its interpretation in terms of my article. Then I will address some of the issues raised by each discussant.

1. A MATHEMATICAL STATEMENT OF RUBIN'S MODEL

In its simplest form, stripped of all of the interpretative language, Rubin's model is a quadruple, $R = (U, K, Y, S)$, in which U and K are sets, Y is a real-valued function defined on $U \times K$, and S is a mapping from U to K. In the language of the article the meaning of the components of R is as follows. U is the population of units, and K is a set of labels or descriptions of the various causes or treatments under consideration. For any $u \in U$ and $k \in K$, $Y(u, k)$ is the value of the response that would be measured on u if u were exposed to cause k. The value of $S(u)$ is the cause or treatment to which u is actually exposed prior to the measurement of the response. In the article I used the equivalent subscript notation, that is, $Y_k(u) = Y(u, k)$, and I let $K = \{t, c\}$. Of course, in general K could contain more than two elements.

In real applications of Rubin's model other measurements besides the response Y need to be represented. I think that all measurements should be regarded as functions defined on $U \times K$, just as Y is. If X is such a function, then $X(u, k)$ is the value of the X measurement that would be made on u if u were exposed to cause $k \in K$. One special type of measurement needs mention here. If the value of $X(u, k)$ does not depend on which cause k to which u is exposed I shall call X an *attribute* of u; that is, $X(u, k) = X(u)$ for all $u \in U$ and $k \in K$. Important examples of attributes are (a) pre-exposure variables (Sec. 3) and (b) post-exposure variables that cannot be affected by k. Among the measurements that are *not* attributes I include other response variables besides Y and "post-treatment concomitant variables" (Rosenbaum 1984b).

The purpose of Rubin's model is to provide a language for discussing causation, and this language takes *units*, *causes*, and *responses* as primitive notions that are not defined further. These three elements, however, are not arbitrary and must satisfy the basic property that Y is defined on all of $U \times K$. The *effect* of cause t relative to c is then defined in terms of these primitive notions, that is, as $Y(u, t) - Y(u, c)$, and the observed response on each unit is also defined in terms of the elements of R, that is, $Y_S(u) = Y(u, S(u))$.

By taking units, causes, and responses as the primitives of his theory and defining effects and observed data in terms of them, Rubin's model breaks with an ancient philosophical tradition that takes "events" or "phenomena" as primitives and attempts to define what is meant by one event being *the* cause of another.

An *application* of Rubin's model requires an identification of the elements of R with features of a real-world problem. What are the units, the causes, the responses? How are units actually exposed to the action of the causes? Is Y defined on all of $U \times K$? If the identification of the elements of the real-world application with those of Rubin's model leads to a faithful representation of the real-

Journal of the American Statistical Association
December 1986, Vol. 81, No. 396, Theory and Methods

world situation by the model, then it becomes a useful framework for making statements about cause and effect. If the representation is not faithful, then Rubin's model does not apply and cannot be used to make causal statements. The question of the "faithfulness" of a particular representation is, in my opinion, one on which people may disagree. For example, it is usually easy to make an identification of U, K, and Y with units, treatments, and a response variable in a randomized experiment, but complex observational studies can provide cases in which reasonable people might disagree as to the proper identification of the elements of the model.

2. RUBIN'S COMMENTS

What Rubin's SUTVA lacks in mellifluence it more than makes up for in utility. I view the SUTVA as a general purpose way of checking on the faithfulness of a particular specification of Rubin's model as a representation of a real-world application. Rubin's comments on SUTVA are illuminating and I would like to add a few of my own.

One might wonder how Hume's notion of "temporal succession" fits into the abstract formulation of Rubin's model given earlier, which does not involve time, explicitly. I view temporal succession as a part of the *application* of the model rather than as a part of the model itself. For example, the value of $Y(u, k)$ is supposed to depend on u and k. For this to happen the exposure of u to k must occur prior to the measurement of $Y(u, k)$. This *forces* temporal succession upon us. Under SUTVA the value of $Y(u, k)$ depends on (u, k) but *does not depend on anything else.* SUTVA and temporal succession are, therefore, two sides of the same coin. As Rubin points out, Fisher's null hypothesis corresponds to $Y(u, k) = Y(u)$ for all $k \in K$. I will point out that unit homogeneity (Sec. 4.2) corresponds to the parallel assumption that $Y(u, k) = Y(k)$ for all $u \in U$. Both Fisher's null hypothesis and unit homogeneity are special cases of SUTVA.

Rubin is correct in pointing out that I view as meaningless "causal" statements in which the "cause" is an attribute of the units. By this I simply mean that causal effects are not well defined in such cases, because Y is not defined on all of $U \times K$, as I discuss in Section 9. Rubin accepts such statements as meaningful in circumstances when they can be construed as rejections of Fisher's null hypothesis that are made without clear statements as to what c is or what $Y_c(u)$ is. Glymour also objects to my use of the term "meaningless" on more general grounds. Rubin and Glymour may be right, but I would call such statements "causally innocuous," since they are of such a general nature as to have no useful consequence in the real world.

Rubin's analysis of Neyman's null hypothesis is illustrative of the value of Rubin's model. By using the SUTVA, Rubin gives meaning to Neyman's notion of "technical errors," which I ignored in my analysis. I ignored technical errors because I find their source of probability to be completely artificial. For example, Neyman (1935) described the source of probability in this way.

> Suppose that we repeat the experiment indefinitely without any change in vegetative conditions or of arrangement so that the kth object is always tested in plot (i, j). The yields from this plot will form a population, say $\pi_{ij}(k)$ and $X_{ij}(k)$ will be defined as the mean of this population. (p. 110)

In fact, we cannot perform such an experiment over and over again, so what did Neyman really intend? I think that Rubin's analysis is very neat and that it does give a meaning to Neyman's technical errors that is easy to understand and that can lead to interesting statistical analyses.

Nevertheless, I think that the problem Neyman and Fisher were addressing does not depend on the existence of technical errors and would still be there if SUTVA were satisfied with only two causes in K (as I assumed). Readers will have to judge for themselves which analysis they prefer, but I encourage Rubin to provide us with a full-blown analysis of the Latin square along the lines indicated in his discussion, as this may add another interesting chapter to this classic problem.

3. COX'S COMMENTS

It was a relief to find that Cox agrees with me that "certain variables cannot properly be regarded as causes." After reading the comments of the other discussants I was beginning to wonder if this view, which I regard as perfectly obvious, was shared by no one else.

I think Cox's term *intrinsic variable* is what I have meant by *attribute* in this rejoinder. Intrinsic variables that are "associated with the environment" can be competing, uncontrolled causes, but I do not believe that they need to be treated as such in the analysis of experiments or observational studies. After all, rainfall and soil fertility may be associated with each other in complicated ways, but it is possibly best to regard them as attributes of a given field over a given time period.

Cox raises what I regard as a very important point about "unit-treatment additivity" or, as I prefer to call it, the assumption of constant effect (Sec. 4.4). If there are no other measured variables besides Y, then it is impossible to falsify the constant effect assumption *with the data in hand.* This is true regardless of the sample size. When there is an attribute or intrinsic variable $\mathbf{X}$ on the scene, then we may be able to falsify the constant effect assumption, but we cannot falsify the *conditional* constant effect assumption that holds conditionally for each value of $\mathbf{X}$, that is,

$$T(\mathbf{x}) = Y_t(u) - Y_c(u) \qquad \text{for all } u \in U \text{ such that } \mathbf{X}(u) = \mathbf{x}.$$

It is natural to consider applying Occam's razor to such situations and to make the appropriate (conditional) constant effect assumption the starting point for analyses of such data. Such a view makes one sympathetic with Fisher's side of the Fisher/Neyman argument described in Section 6, in my opinion.

4. GLYMOUR'S COMMENTS

I am extremely grateful for Glymour's willingness to bring a philosopher's point of view to this discussion. Rubin and I have always been aware of the "subjunctive" quality of the definition of a causal effect—the "woulds," "ifs," and "weres" of that definition—but I was not aware of the relevance of counterfactual conditionals until I read Glymour's comments on my article. I especially like the notion of "possible worlds," since this is what I think the function Y is intended to represent. For unit u, $Y(u, \cdot)$ represents all of the relevant possible worlds for u. On the other hand, $S(u)$ and $Y(u, S(u))$ described the world that actually exists (for observational studies) or the world that will be observed (for experiments) for unit u.

I must disagree with Glymour's paraphrasing of my (i.e., Rubin's) analysis, however, and with the counterfactual analysis of causation of Lewis described by Glymour. I believe that there is an unbridgeable gulf between Rubin's model and Lewis's analysis. Both wish to give meaning to the phrase "A causes B." Lewis does this by interpreting "A causes B" as "A is *a* cause of B." Rubin's model interprets "A causes B" as "the *effect* of A is B." Rubin adopts the notion of an experiment as the fundamental way of thinking about causation, studying the effects of known causes. Lewis starts with the effect and, like Aristotle, seeks to define what it means to be *a cause* of that effect. Can Rubin's model ever define what it means for "A to be *a* cause of B"? I do not think so. In Section 9, I convinced myself, at least, that statements like "A is *a* cause of B" are generally false and always depend on our current state of knowledge. Notice that once a statement of the form "the *effect* of A is B" has been experimentally verified it does not go away or become false as our knowledge of the subject increases. Old, replicable experiments never die, they just get reinterpreted.

I think that Glymour's criticisms are more directed at the way in which Rubin's model might be applied than at the model itself. For example, he interprets my use of "attribute" to refer to such things as "genetic constitution" and then points out that we might be willing to "identify persons across . . . some alterations in genetic structure." Such identification would produce an attribute that is a cause. I would see it differently. As technology evolves so do the types of causes or treatments that can be applied. This is the history of medicine, for example. The units must change, of course. In the Down's syndrome example, they change from a person to a zygote. What was an attribute at one level could be manipulated at a different level of analysis.

5. GRANGER'S COMMENTS

I agree with Granger that statisticians are all too willing to shirk the responsibility of addressing issues of causality. His work in this area captures aspects of causation that many find attractive—compare Sections 5.3 and 8.2. His point of view is quite different from that discussed in the article. To illustrate this consider Granger's example of a cross-sectional causal question, for example, "why does one household spend more on electricity . . . than does another." This is a comparison of the responses of two distinct units, for example, $Y_S(u_1)$ versus $Y_S(u_2)$, rather than a comparison of the responses of a unit under two causes, for example, $Y_t(u_1)$ versus $Y_c(u_1)$. I regard the values of $Y_S(u_1)$ and $Y_S(u_2)$ to be of *no causal interest* unless they shed light on the value of a causal effect such as $Y_t(u_1) - Y_c(u_1)$. The Fundamental Problem of Causal Inference (Sec. 3) must be faced and overcome in some way so that the data values $Y_S(u_1)$ and $Y_S(u_2)$ can answer causal questions. By focusing on the observed data Granger overlooks what I regard as *real* causal questions that must, of necessity, be couched in terms of information, of which only some can be observed.

I agree with Granger that experiments are not always possible to do in many branches of science and that even when they are, they may not actually answer the questions of interest—note Cox's fertilizer/bird example in this regard. I disagree, however, with the implication that the experimental (i.e., Rubin's) model tells us nothing about nonexperimental causal research. In my opinion, there is no difference in the conceptual framework that underlies both experiments and observational studies—Rubin's model is such a framework. In observational studies we know less about the situation than we do in experimental studies and this lack of information simply serves to make causal inferences from observational studies more speculative than they are in experiments.

Granger expresses the view that the experimental model is not helpful in problems of "temporal causality," which he defines as "causal questions about data that are a group of time series." The idea that Rubin's model is somehow incapable of accommodating time-series data is misleading. There is no reason why the response Y cannot be a function of time rather than simply a real number. Thus $Y(u, k, \alpha)$ is the value of the response that would be measured at time α on unit u if u were exposed to k. The observed data are $Y(u, S(u), \alpha)$ for all relevant α values. Causal effects are more complex than before, since they now involve comparisons of functions, that is, $Y(u, t, \cdot)$ and $Y(u, c, \cdot)$. This might be done with functionals that associate single numbers with $Y(u, t, \cdot)$. More complicated issues arise if the causes of interest are themelves functions of time; that is, K is a set of functions and $k(\alpha) \in K$ describes the "level" of a cause at time α. These added complexities have not been analyzed carefully as far as I know and ought to be pursued to clarify the problem of causal inference in a time-series setting. Careful attention to Hume's "temporal succession" is critical in such settings.

Finally, I must strongly disagree with Granger's (and I believe Glymour's) view that, for example, questions such as "race . . . affects . . . crime rates" and "the death sentence cause(s) decreases in murder rates" are on the same causal footing. In the former, "race" cannot be manipulated, whereas in the latter "the death sentence" is manipulated by governors and legislators all the time. The former is an associational statement that is not uninflammatory, and the latter is a causal statement of great public policy interest—regardless of how well or poorly it may have been studied by enthusiastic regression modelers. Granger's theory of temporal causality as expressed in Section 8.2 and in his comments contains, in my view, too generous a definition of causality. I find it, at bottom, indistinguishable from association.