STATISTICS IN MEDICINE, VOL. 12, 1477–1494 (1993)

CLINICAL TRIALS AND CAUSATION: BAYESIAN PERSPECTIVES

KENNETH F. SCHAFFNER

The George Washington University and the University of Pittsburgh, 714 T Gelman Library, George Washington University, Washington, DC 20052, U.S.A.

SUMMARY

In addition to the *safety*, it is essential to establish the *causal efficacy* of extant and new treatments, and well-designed clinical trials are thought by most to be the 'gold standard' to accomplish this. Contrary to most statisticians' and regulators' views, however, I will argue that the *concept* of causation involved in clinical trials is not all that clear. I discuss the manipulability approach to causation, interpreted counterfactually, which seems to fit causation as it is found in such sciences as physiology, but it has unclear relations to a concept of causation proposed by a number of epidemiologists. I characterize 'epidemiological causation' as probabilistic and formulated at a population level, and dependent on certain general criteria for causation as well as study-design considerations. I then attempt to clarify the connections between these concepts of causation and Cartwright's views on complexity and causality, a 'Bayesian' framework proposed by Rubin and further elaborated by Holland, and Glymour and his colleagues' recent directed graphical causal modelling approach.

1. INTRODUCTION

In addition to the *safety*, it is essential to establish the *causal efficacy* of extant and new treatments, and well-designed clinical trials are thought by most to be the 'gold standard' to accomplish this. However, the *concept* of causation that is involved in clinical trials is not, on reflection, all that clear. This claim about the deeper level opacity of the notion of causation in clinical trials is probably contrary to most scientists' views, including both statisticians and regulators; thus I will need to provide some support for this claim. In this paper I will maintain that there is a basic concept of causation that can be made clear which holds for biological and medically relevant *mechanisms*, but that clinical trials rarely (if ever) capture that sense of causation; instead they involve a derivative but still very important concept of *epidemiological causation* or *clinical causation*. I also will argue that non-experimental study designs such as historical studies also involve clinical causation, but that because of the chance for additional sources of bias, these study designs are less reliable than the randomized controlled clinical trial.

This paper will make use of what several epidemiologists have written on causation, and I will, in the later portions of it, also turn to some insightful approaches to causation relevant to clinical trials that have been recently advanced by Nancy Cartwright[1] and by Clark Glymour and his colleagues.[2,3] I will develop my presentation by discussing one Bayesian perspective on clinical trials, specifically a thesis initially put forth in 1978 by Rubin[4] and subsequently extended by him,[5] as well as by Pratt and Schlaifer[6,7] and Holland.[8] I interpret Rubin's framework as a statistical implementation of what was developed later in a more philosophical context by Cartwright.[1] In addition, Glymour and his colleagues[3] have very recently taken Rubin's framework as well as

0277–6715/93/161477–18$14.00

Pratt and Schlaifer's extensions of that framework, and unified Rubin's approach with a directed graphical causal modelling approach. I believe this unification offers much promise for probing the foundations of what I term clinical causation.

2. THE MANIPULABILITY/COUNTERFACTUAL APPROACH TO CAUSATION

A review of the various approaches to causation that have been taken by philosophers over the past two millennia suggests that the concept of causation may not be unitary. In Aristotle we find four different senses of the term, and post-Humean analyses comprise such diverse approaches as regularity and conditional accounts, the activity or manipulability view, the (rationalist) logical entailment theory, a non-logical entailment version, and the more recent possible world accounts.[9-11] In my view, several of these diverse approaches need to be drawn on and intertwined to constitute an adequately robust analysis of causation for biology and medicine. I will not have an opportunity to do that in this article (I have done so elsewhere[12]), but I believe that I can make my points sufficiently well by referring to one such approach to causation which is widespread among statisticians, including I think those involved with clinical trials. This is the 'manipulability' approach to causation. It can be found expressed in the second edition of the widely used *Statistical Package for the Social Sciences* or *SPSS* manual[13] in the following manner:

> We propose the following 'operational' definition as an initial approximation to the idea of causation: X_1 is a cause of X_0 if and only if X_0 can be changed by *manipulating* X_1 and X_1 *alone*. We note first that the notion of causation implies prediction but prediction of a particular kind. It implies the notion of *possible* manipulation.[13]

In addition, the *SPSS* manual adds:

> The preceding definition of causation suggests both the criterion of causation and the means to measure causal effects. First to establish conclusively that X_1 is a cause of X_0, one must perform an 'ideal' experiment in which all the other *relevant* variables are held constant while the causal variable is being manipulated. Second there should be some accompanying change in the dependent variable. We will use such validation as the ultimate criterion that X_1 is the cause of X_0.[13]

I cite this characterization from the *SPSS* manual not because this is an authoritative source for either the statistical or philosophical communities, but rather to indicate the widespread acceptance of this manipulation approach to causation. This approach, which is also sometimes referred to as the 'activity theory', can be found in the earlier work of Hart and Honoré,[14] as well as in the work of Collingwood,[15] Gasking[16] and von Wright;[17] it is also stressed in Holland[8] and Rubin.[5] It has been criticized as not being able to distinguish causality from mere correlation, but it fails on this ground only if the *counterfactual* interpretation of this approach is disallowed (also see Holland[8] and Glymour[18]). The notion of 'counterfactual' is a pervasive one in philosophy and is of especially critical importance in discussions of causation. (Essentially the term refers to a state of affairs which is literally false, but on the assumption of a *factually different* antecedent condition, *would be* the case.) In their forthcoming book *Causation, Prediction, and Search*, Spirtes *et al.*[3] also defend a manipulability approach, but one which explicitly permits counterfactuals. This counterfactual component is also contained in the earlier quote from the *SPSS* manual, where the notion of a *possible* manipulation was explicitly appealed to.

That this approach may also be the sense of causation involved in clinical trials receives support from Howson and Urbach,[19] who point out after describing a comparatively simple clinical trial with a control group that:

> the reason for a control group is obvious. One is interested in the *causal effect* of the drug on the chance of recovery; so ideally one wants to compare how patients responded with the drug with how they *would have responded* without it. (my emphasis)

Again it is important to note here that the idea of producing one thing by doing another has had a *counterfactual* aspect added to it. Other philosophers have made a similar point. For example Mackie[20] defends the following view: 'the distinguishing feature of causal sequence is the conjunction of necessity-in-the-circumstances with causal priority.' What this means is that '*X* is necessary in the circumstances for and causally prior to *Y* provided that *if X were kept out of the world* in the circumstances referred to and the world ran on from there, *Y* would not occur' (my emphasis).

3. THE EPIDEMIOLOGICAL NOTION OF CAUSATION: PROBABILISTIC CAUSATION?

The epidemiologists' notion of cause

Several different groups of epidemiologists such as MacMahon and Pugh,[21] Fletcher *et al.*[22] and also Kleinbaum *et al.*[23] have put forth the interesting idea that epidemiologists use a different notion of cause than do, for example, physiologists and molecular biologists, whose notion might appropriately be termed 'physiological causation' or perhaps, in connection with diseases, 'pathophysiological causation'. Though I use the term 'epidemiological' in this section and throughout this article, I view the randomized controlled clinical trial as falling within the scope of the epidemiological study, albeit in a form which is taken to be the best of such epidemiological studies in terms of strength and reduction of bias. Epidemiologists MacMahon and Pugh suggested that 'a causal association may usefully be defined as an association between categories or events or characteristics in which an alteration in the frequency or quality of one category is followed by a change in the other.'[21] Fletcher *et al.* similarly wrote: 'when biomedical scientists study causes of disease, they usually search for the underlying pathogenetic mechanism or final common pathway of disease.'[22] Though these authors indicate their agreement with the importance of this approach to causation, they add that:

> The occurrence of disease is also determined by less specific, more remote causes such as genetic, environmental, or behavioral factors, which occur earlier in the chain of events leading to a disease. These are sometimes referred to as 'origins' of disease and are more likely to be investigated by epidemiologists. These less specific and more remote causes of disease are the risk factors.[22]

Finally let me cite Kleinbaum *et al.*, who write:

> In epidemiology, we use a probabilistic framework to assess evidence regarding causality – or more properly to make causal inferences. . . . But a probabilistic viewpoint does not automatically negate our belief in a (modified) deterministic world. . . . In other words, we need not regard the occurrence of disease as a random process; we employ probabilistic considerations to express our ignorance of the causal process and how to observe it.
>
> Because of the lack of certainty in our results, epidemiologists generally use the term risk factor instead of cause to indicate a variable that is believed to be related to the

probability of an individual's developing the disease prior to the point of irreversibility.[23]

These quotations raise the provocative thesis that the notion of causation may be in some interesting methodological and/or substantive sense different for the epidemiologist. If this point of view is correct and can be developed further, it may help us make sense of some of the epidemiologists' other comments concerning the criteria by which they assess causal claims, and assist us in sharpening our analysis of inference from epidemiological studies. In order to answer this question, it will be necessary to examine a notion of causation developed relatively recently by several philosophers known as 'probabilistic causation'. With the aid of this notion, I believe we can make some progress in clarifying the interesting differences between the way in which epidemiologists as opposed to other biomedical scientists approach the notion of cause.

Probabilistic causation

The notion of epidemiological causation appears to me to be superficially similar, though I shall argue in an important sense not identical, to what philosophers have recently termed 'probabilistic causation'. The pioneer in this area is Hans Reichenbach,[24] and other philosophers of science such as Good[25] and Suppes[26] have developed similar ideas; still others have provided extensive criticism of such concepts, among them Salmon.[27] Suppes's approach to probabilistic causation is perhaps the most widely known, but my approach will begin from some useful suggestions made by Giere[28] concerning the relationship between determinism and probabilistic causation in populations. I then examine some of Cartwright's[1,29] views on Simpson's paradox and its resolution. In particular I want to argue that it is the feature of non-identical individuals in populations that is encountered as part of epidemiological research and inference, including clinical trials, that introduces the key difference into the account of causation ascribable to such systems.

There are some useful distinctions which appear in Giere's account which are not explicit in most discussions of probabilistic causation with which I am familiar. First, as I shall describe in detail in the next paragraph, Giere differentiates between deterministic and stochastic systems, but permits *both* types to exhibit probabilistic causal relationships. This is important since it distinguishes two different ways in which determinism may fail to hold for populations (see below). This is useful for my account because it licenses a more coherent fit between causation in physiologically characterized biological systems and those which we approach from an epidemiological perspective. I shall develop the thesis below that both epidemiologically studied and physiologically studied systems can be deterministic, but that the former will display a probabilistic type of causation best characterized at the population level. This thesis does not, however, exclude stochastic components in physiological processes, but I will argue that in this domain a stronger condition of homogeneity is satisfied that is not met in epidemiological systems. Giere also invokes Popper's 'propensity' interpretation of probability for stochastic systems, and in particular maintains that this gives a means of directly attributing probabilistic causality to individuals rather than to populations. Howson and Urbach[19] define this propensity interpretation as follows: 'certain types of repeatable experiments are endowed with dispositions or *propensities* to produce fixed limiting frequencies of their various outcomes were they to be continued indefinitely under similar conditions.' Now in my view the propensity interpretation is needed only for irreducible singular non-deterministic causation, which, though it may have much to recommend it in quantum mechanical situations, is less clearly demanded in biomedical contexts, though it is possible that those types of studies involving repeated trials on an individual (termed $N = 1$ studies) might favour this interpretation. I prefer a frequency interpretation of

probability which I maintain is more natural in epidemiological contexts where application of epidemiological conclusions to individuals is concerned, and though I will return to this issue later, it is not a point I can pursue in any depth in the present paper.

In Giere's analysis of deterministic systems, a causal relation between C and E is not necessarily a universal relation. An example which Giere often uses is smoking and lung cancer. This is a very useful example for my purposes since the evidence for such a claim is largely epidemiological, but the same general point could be made using any one of the many studies of risk factors for diseases. The entities in Giere's account are individual deterministic systems which can differ in their constitution, so that different individuals with the same causal input C (e.g. smokers) may or may not exhibit E (e.g. lung cancer). Furthermore, since E may come about from a different cause than C, some individuals may exhibit E but not have had C as a causal input.

For a population of deterministic systems, some number of individuals with input C will manifest E and some will not. On Giere's deterministic approach this is because, for any given individual in the population, a universal law $L(C) = E$ is either true or false, depending on that individual's constitution. (I am assuming that this approach to an individual will still permit the individual to exhibit varying outcomes at different times, as in an $N = 1$ study, since the circumstances affecting that individual could alter over time.) An actual population can be examined to yield a relative frequency $\#E/N$, where N is the number of the individuals in the population and $\#E$ is the number of individuals exhibiting effect E. Thus for a population with input C the probability that N individuals will exhibit E will be $P_C(E) = \#E/N$. Likewise for a population without C as input – where $\sim C$ is the input – the probability that N individuals will exhibit E will be $P_{\sim C}(E) = \#E/N$, where this $\#E$ may differ from the previous $\#E$. For Giere, then, this fraction $\#E/N$ has the properties of a probability. (It might be noted, however, that Cartwright's view, discussed later in this article, differs, arguing that frequencies are not, without making some additional assumptions, probabilities.)

Giere prefers to use idealized counterfactual populations for which outcomes are well defined, and he argues that this idea is what one finds in the typical randomized clinical trial, a point with which I shall disagree further below. Thus, by hypothesis, two counterfactual populations which are counterfactual counterparts of an actual population of interest are envisaged, and one is provided with causal factor input C and the other with input $\sim C$. Each counterfactual population will exhibit some number of effects $\#E$ which will be less than its N. Then Giere[28] defines a positive causal factor as: C is a positive causal factor for E in [population] U if and only if

$$P_C(E) > P_{\sim C}(E).$$

A reversed inequality will yield a definition for a negative causal factor, and equality will indicate causal irrelevance. (Note that this notion is almost identical to the definition for a prima facie cause in Suppes's[26] system. What is different is the interpretation in the context of explicitly deterministic systems, and an explicit counterfactual account.)

Giere also introduces a measure of effectiveness of C for E in population U, namely:

$$Ef(C, E) =_{Df.} P_C(E) - P_{\sim C}(E).$$

Interestingly, the measure of effectiveness introduced here is essentially identical to what the epidemiologists have termed a 'measure of effect' for 'attributable risk'. Giere does not comment on this notion, but see Fletcher *et al.*[22] In my view this account of Giere's indicates how systems that are deterministic at the level of the individual can nevertheless exhibit a type of 'probabilistic causation' that bears striking analogies to the concept of causation defended by the epidemiologists quoted earlier. Giere does not limit his analysis to such deterministic systems, however, and does extend his account to cover stochastic systems,[28] but it will not be necessary in this paper to

discuss this extension in any detail. Suffice it to note that the extension to stochastic systems involves the propensity interpretation of probability discussed earlier, and that such an extension could characterize irreducibly probabilistic systems, such as those encountered in quantum mechanics.

As I have noted above, Giere prefers to use counterfactual populations for which inputs are well defined, and he argues that this situation is what one finds in the typical randomized clinical trial. I think this is a mistake on Giere's part, since such a trial makes use of relevantly similar *actual* populations for which it *hopes* to control interfering factors (1) by matching and stratification if the factors are known or suspected to operate, and (2) by randomization for all other unknown interfering factors. Thus a clinical trial hopes at best only to approximate partially the major feature of such counterfactual test populations, in that the investigator hopes to have *only* specified and testable relevant differences between experimental and control groups. The appeal to counterfactuality, however, *is* essential if Simpson's paradox, to be considered in the next paragraph, is to be avoided. I also believe it is precisely this appeal to a counterfactual situation that points toward the difference between the epidemiologists' notion of causation and the probabilistic causal notion, though in Giere's account this distinction is left largely implicit. The counterfactual interpretation of causation is also what we find at the basis of what I term physiological or scientific causation in this paper; thus both physiological causation and probabilistic causation is distinguished from epidemiological causation. This distinction is presented in an explicit manner in Cartwright's analysis, to which I now turn.

Several essays and books by Cartwright[1,29,30] have focused increased attention on the counterfactual foundation underlying a probabilistic conception of causality. Cartwright[29] begins by citing a paradox known as Simpson's paradox (or the Yule–Cohen–Nagel–Simpson paradox, to give all who have identified it their due). This paradox shows that any association which holds between two variables in a population which can be used to license a relation of probabilistic causation of the type characterized in Reichenbach's and Suppes's approaches 'can be reversed in the subpopulations [of that population] by finding a third variable which is correlated with both'.[29] Her example is relevant to our discussion. Suppose that cigarette smoking is a weak cause, in the sense of a risk factor, for myocardial infarctions. But suppose that, in the population examined, cigarette smoking is associated with physical exercise, and also suppose that such exercise is strongly preventive of infarctions. Then unless subpopulations are examined in which the exercise factor is held fixed, cigarette smoking will be concluded to be a *preventive* factor.

Cartwright's solution for Simpson's paradox is to appeal to a counterfactual condition. Simply put, she writes: 'C causes E if and only if C increases the probability of E in every situation which is otherwise causally homogeneous with respect to E'.[29] (I think that Cartwright's definition of cause here is essentially identical to Giere's introduced above, with the somewhat superficial difference that Cartwright formally includes homogeneity in her definition whereas Giere introduces the idea implicitly into his counterfactual interpretation.) This notion of a causally homogeneous situation has subsequently been re-evaluated and generalized, and also explored for its important implications for singular causation in her most recent book.[1]

In a formalism which perhaps indicates more explicitly the nature of the counterfactual conditions underlying Cartwright's definition of causality, she writes that:

> to test for a causal connection between a putative cause C and an effect E, it is not enough to compare $P(E|C)$ with $P(E|-C)$. Rather one must compare $P(E|C \pm F_1 \ldots \pm F_n)$ with $P(E|-C \pm F_1 \ldots \pm F_n)$ for each of the possible arrangements of E's other causes, here designated by $F_1 \ldots, F_n$. The symbol $\pm F_n$ indicates a definite choice of either F_n or $-F_n$.[1]

These test factors whose presence and absence in addition to C is presupposed would block problems associated with Simpson's paradox. Cartwright adds that this approach suggests a new (but generally equivalent to her earlier[29]) form of what she terms her CC principle:[1]

$$\text{CC: } C \text{ causes } E \text{ iff } P(E|C \pm F_1 \ldots \pm F_n) > P(E| - C \pm F_1 \ldots \pm F_n),$$

where $\{ \pm F_1 \ldots \pm F_n, C\}$ is a complete causal set for E. (Exactly what factors can be allowed into such a formulation is somewhat complex. In his book[31] Sober points out that unless some restrictions are imposed, probabilities may become ill-defined; also see Eells and Sober.[32] In addition, the factors may need to be temporally indexed.)

I believe that this approach to blocking the paradoxical effects of selecting subpopulations suggests the root of the difference which the epidemiologists perceive between their concept of causation as cited above and what might be termed the physiological or scientific (in the sense of laboratory or bench science) notion of causation. I believe that the reason is not that epidemiological research involves probabilistic elements, but rather that epidemiology as practised cannot be certain that it has examined homogeneous populations, or even sufficiently homogeneous populations such that a treatment effect will be identical for all experimental subjects. Physiological causation of the type employed by molecular biologists, say, deals with idealized models and often experiments with specially bred strains of organisms, and can thus be reasonably certain that its populations are homogeneous and that equal causes will yield equal effects. I will not be able in the context of this paper to develop extensively detailed arguments to support this thesis, and must refer the reader to a recent essay of mine[33] for such argumentation. Because this thesis is central to the present paper, however, and because I believe that sketching the difference between what I term physiological causation and epidemiological or clinical causation is supported by an elaboration of Cartwright's views cited above, I want to briefly outline this notion of physiological causation. I will begin with two simple examples from muscle physiology. I will then explore similar ideas in the statistical literature where such notions are evoked by the phrases 'unit homogeneity' and 'assumption of constant effect', the latter also being known as 'additivity'.

4. PHYSIOLOGICAL CAUSATION IN A COMPLEX WORLD: THE UTILITY OF LEVELS IN ACHIEVING SIMPLIFIED CAUSAL GENERALIZATIONS

Beginning biology students learn about many different types of physiological mechanisms. In their first introductions to the subject the mechanisms are idealized and simplified. As they become more knowledgeable they learn about the many mutations and variations from the simplified mechanisms that exist. If they embark upon research, they also learn to use these variations as subtle applications of Mill's method of difference or Claude Bernard's method of comparative experimentation to license causal claims about living organisms' processes.[34] In several other publications I have developed the thesis that what function as a surrogate for *theories* in much of biology, including molecular biology, are *families* of similarly related models, typically of an interlevel character, operating according to generalizations of both broad and narrow scopes.[12,35,36] These generalizations constitute the 'laws of working', as it were, of fundamental biological mechanisms, but the picture I present in these recent publications emphasizes the variation, diversity, and complexity found in fundamental mechanisms. These mechanisms may be either deterministic or stochastic. Let me refer to one example of each type quite briefly, using very simple accounts which may be slightly misleading in relation to the actual complexity of the mechanisms, but which will, I think, make the point I want to here about physiological causation.

My first exemplar is drawn from muscle physiology.[37] We investigate the cause of the increased force of contraction which skeletal muscle delivers when stretched prior to the administration of a contracting impulse when compared with the same muscle in an unstretched state. We determine that the increased force is a consequence of additional overlap, generated by the stretch of the muscle, between the two sets of muscle filaments known as actin and myosin comprising the sarcomere (filament bundle), since this additional overlap increases the number of reactive sites where chemical energy is transformed into the energy of motion. This is known as the sliding filament model of muscle physiology.

The second example comes from neurophysiology. We want an account of how a nerve can stimulate a muscle on which it impinges to contract. It is found that the signal for contraction is carried by chemical neurotransmitter molecules that are probabilistically released from the nerve endings in packets or 'quanta', and that these quanta diffuse across the neuromuscular junction space and cause the opening of microscopic channels on the muscular side, resulting in an electrical signal that causes muscle contraction. Furthermore, 'fluctuations in release [of the quanta] from trial to trial can be accounted for by binomial statistics . . . [and] when the release probability p is small . . . the Poisson distribution provides a good description of the fluctuations.' This tells us that 'release of individual quanta from the nerve terminal . . . [is] similar to shaking marbles out of a box' through a small hole.[38] This model of a probabilistic process yields excellent agreement with what is observed in a variety of micro-experiments at the neuromuscular junction involving neural signals causing muscular contractions.

Both of these models have been extraordinarily well confirmed, and in the form I present them satisfy a strong condition of homogeneity. In these models all myofilaments are identical and all neurotransmitter quanta are essentially identical. As molecular biology has progressed, scientists can now identify genes that are responsible for specific and detailed components of mechanisms such as described above, and can clone these genes, sequence them, and identify (or create if necessary) new genes with subtle differences that can be used as contrast cases to characterize the action of such mechanisms. At this type of level then one can satisfy Cartwright's CC principle.

However, biologists are not interested in just molecular mechanisms; they often explore higher levels of aggregation, searching for generalizations stateable in language above the molecular level. Exactly how to sort out such levels has been contentious among biologists and philosophers of biology, but I find the suggestion made by Wimsatt[39] a useful one. Wimsatt proposed a kind of pragmatic notion of a level of organization:

> If the entities at a given level are clustered relatively closely together (in terms of size, or some other generalized distance measure in a phase space of their properties) it seems plausible to *characterize a level as a local maximum of predictability and regularity* . . . . [S]upposing that . . . regularity and predictability of interactions is graphed as a function of size of the interacting entity [for example,] . . . [t]he levels appear as periodic peaks, though of course they might differ in height or 'sharpness'.[39] (my emphasis)

Wimsatt provided a graphical representation of this interpretation, and also speculates on other less 'sharp' and less regular possibilities where the utility of level individuation/separation would be much less valuable.[39]

This pragmatic suggestion seems generally correct to me, and also suggests the reason why it is difficult to give any general and abstract definition of a level of aggregation, namely that a level of aggregation is heavily dependent on the generalizations and theories available to the working biomedical scientist. Thus if we were to have no predictive, generalizable knowledge about the properties and functions of ribosomes, a ribosomal level of aggregation would not be one worth characterizing.

The import of Wimsatt's view for my concerns in the present paper is that simplification and reduction of diversity can occur at a variety of levels of aggregation. In order to achieve generalizations which are useful we should take them where we can find them. Underlying diversity can in a wide number of instances be masked by generalizations formulated at higher levels.

In her two books[1,30] Cartwright has stated that 'nature is complex through and through: even at the level of fundamental theory, simplicity is gained only at the cost of misrepresentation.' Cartwright criticizes Glymour *et al.*'s[2] approach to discovering causal regularities from this point of view. But it seems to me that this is too quick from a methodological point of view. Simplicity considerations I would maintain *are* useful in selecting for test and elaboration those generalizations at appropriate levels where generalizations have reasonably broad scope and are not hedged with innumerable qualifications and exceptions. Thus, though I would agree wholeheartedly with Cartwright as a defender of enormous diversity and complexity in nature, I also believe that we could not function either as lay persons or as scientists if we did not seek out and use generalizations that were 'simple' and of broad scope. Furthermore I view these generalizations not as 'misrepresentations' but more as 'approximations' which will hold in a sense 'on the average' or 'for the most part'. This view is, I think, what is at the root of the difference between the type of causation we find exemplified in physiological investigations at the molecular level – the examples of the mechanisms I introduced just above – and the causal claims that epidemiologists make and that clinicians obtain from the results of clinical trials. However, because we cannot satisfy Cartwright's CC condition we must employ *other* means of avoiding spurious causation, which leads me back to the epidemiologists' additional recommendations for arriving at reliable causal claims.

5. THE ROLE OF STUDY DESIGNS AND CRITERIA FOR LICENSING EPIDEMIOLOGICAL OR CLINICAL CAUSATION

I believe it is because epidemiologists, as well as individuals who design and evaluate clinical trials, work with not necessarily homogeneous populations that additional means have been developed to guard against spurious causality attributions arising from such investigations. Thus we can find in standard epidemiological textbooks various guidelines or criteria offered by epidemiologists that allow a discrimination between causal claims and accidental associations. The interesting feature about these criteria is that they have departed significantly from the universalistic deterministic approach represented by Koch's postulates for causation, which I think are their historical protosource, and have become frankly and explicitly probabilistic.

The first set of such epidemiological criteria for causation appeared in the eighth edition of Sir Austin Bradford Hill's *Principles of Medical Statistics* in 1966. In the Preface to that edition, Hill wrote that:

> A . . . development in recent years has been the increase in research into the environmental features associated with chronic diseases with the object of determining factors in their aetiology. This research based upon the epidemiological-statistical approach often raises the difficult problem of distinguishing causation from association. To my final chapter . . . I have added a discussion of this problem.[40]

Kleinbaum *et al.*[23] suggest that it was in no small part the research leading to the 1964 Surgeon General's report on smoking and health that stimulated the epidemiologists to begin a conscious analysis of epidemiological causation. Kleinbaum *et al.* also cite Hill's versions of those criteria, but I think it would be more useful for our purposes to examine similar but somewhat further

developed criteria proposed by Evans[41] and adopted by Lilienfeld and Lilienfeld.[42] These criteria are as follows.

Evans's criteria as modified by Lilienfeld and Lilienfeld

(The terms in brackets [ ] are Evans's original expressions for the similar idea.)

1. Prevalence of the disease should be significantly higher in those exposed to the [putative] hypothesized cause than in controls not so exposed (the cause may be present in the external environment or as a defect in host responses).
2. Exposure to the hypothesized cause should be more frequent [present more commonly] among those with the disease than in controls without the disease – when all other risk factors are held constant.
3. Incidence of the disease should be significantly higher in those exposed to the cause than in those not so exposed, as shown by prospective studies.
4. Temporally, the disease should follow exposure to the hypothesized causal agent with a distribution of incubation periods on a log-normal curve [bell-shaped curve].
5. A spectrum of host responses should follow exposure to the hypothesized agent along a logical biologic gradient from mild to severe.
6. A measurable host response following exposure to the hypothesized cause should have a high probability of appearing [should regularly appear] in those lacking this before exposure (for example, antibody, cancer cells) or should increase in magnitude if present before exposure; this response pattern should occur infrequently [should not occur] in persons not so exposed.
7. Experimental reproduction of the disease should occur more frequently [in higher incidence] in animals or man appropriately exposed to the hypothesized cause than in those not so exposed; this exposure may be deliberate in volunteers, experimentally induced in the laboratory, or demonstrated in controlled regulation of natural exposure.
8. Elimination or modification of the hypothesized cause or of the vector carrying it should decrease the incidence of the disease (for example, control of polluted water, removal of tar from cigarettes).
9. Prevention or modification of the host's response on exposure to the hypothesized cause should decrease or eliminate the disease (for example, immunization, drugs to lower cholesterol, specific lymphocyte transfer factor in cancer).
10. All of the relationships and findings should make biologic and epidemiologic sense.

There has been a good deal of contentious discussion in the literature about the value of these guidelines, particularly as applied to the smoking/lung-cancer studies, and Glymour and his colleagues have in a detailed summary of that case described the earlier statement of the Surgeon General's 'epidemiological criteria for causality' as an 'intellectual disgrace' because of their vagueness.

It is of interest to note that the above criteria do not *per se* refer to the nature of the *type of epidemiological study*, even though Lilienfeld and Lilienfeld do discuss the different types and their respective strengths in their book.[42] I tend to interpret this partly on historical grounds and partly on the basis of two somewhat different orientations one can find among epidemiologists. On historical grounds, one can account for the above criteriological approach as the development of a line of thinking from Koch through Hill to Evans (and the Lilienfelds). There exists a somewhat different though not in any sense opposed approach to scrutinizing the nature of causal claims in epidemiology, however, and it places considerably more emphasis on study type and design, appealing to the above quoted criteria in a more secondary sense.

A good example of this second approach to causation is represented by Fletcher *et al.*[22] Their account urges we consider the different types of study designs in evaluating whether causal claims are warranted or not. The design-based approach to evaluating epidemiological causation appears to take its theme from the manner in which prima facie but spurious causal associations might arise, and provides an account based on controlling for the erroneous inferences. This approach begins in a sense from the question raised by Fletcher *et al.* about 'possible explanations for clinical observations', say an observation that *A* appears to be associated with *D*. Fletcher *et al.*'s answer is in a later chapter on 'Cause' returned to in the section on 'Establishing cause' with the following type of response: [22]

1. *A* can (always, often, sometimes) cause *D*.
2. There may be some other confounding factor that causes *A* and *D* to 'travel together', that is, a common cause.
3. The association can be due to bias, for example the ways we selected or measured the cases.
4. The association can be due to chance, for example a run of good or bad luck or a set of coincidences results in an association.[22]

In the context of this view of bias and other confounders (and chance) as possible factors in the production of prima facie causal relations, Fletcher *et al.*[22] cite some of the criteria which are found in Evans's and the Lilienfeld's approach above. The proponents of this design-based approach, however, place *considerably* more emphasis on the differential strength of experimental design, characterizing the case control type of study as the weakest and the randomized clinical trial as the strongest. Fletcher *et al.* write in their section on 'Establishing cause' that 'the most important evidence for establishing a cause-and-effect relation is the strength of the research design used to establish the relation.'[22] The ranking of strength of study type is based in part on the likelihood of bias affecting the validity of causal claims based on such studies. To be sure, those authors cited under the 'Criteriological' heading also discuss different study designs and their pitfalls, the primary difference being in the fact that study design does not explicitly figure in the assessment of causal claims, as indicated in Evans's criteria above.

A most suggestive and quite general account within this second design-based family is Feinstein's proposal. This account is developed over a number of chapters in his relatively recent book.[43] Feinstein provides what he terms an 'intellectual model' that may be used 'to evaluate the scientific quality of cause–effect research'. The model is depicted in Figure 1 and specifies along the top of the model the location of potentially 'attendant major problems'. Though attention to these forms of bias and the reduction of them can be viewed as an attempt to provide homogeneity, and thus control for factors that can generate spurious causal claims, the complexity of the influences on human subjects generally precludes this from being an attainable goal, in the sense that Cartwright's CC condition cannot be met.

I believe these considerations suggest that what epidemiological studies, *including well-designed randomized controlled clinical trials*, yield is not causation in the same sense that the physiologists employ it, but rather what might be termed *epidemiological or perhaps clinical causation.* (I use the term 'clinical' causation in addition to 'epidemiological' causation because I view this type of causation as what is determined by *clinical* trials. In addition, this type of causation is also what the *clinician* encounters in many of her diagnostic, prognostic, and therapeutic interactions with patients.) This type of causation works at the *population level*—at the level of the individual the causation is diffuse and typically inaccessible. A generalization associating a cause with an effect based on populations which have varying characteristics, in situations of gappy knowledge, and in which the association is not universal (either on the basis of variation of initial conditions or on purely stochastic grounds), is a claim about population causality in which causal connections

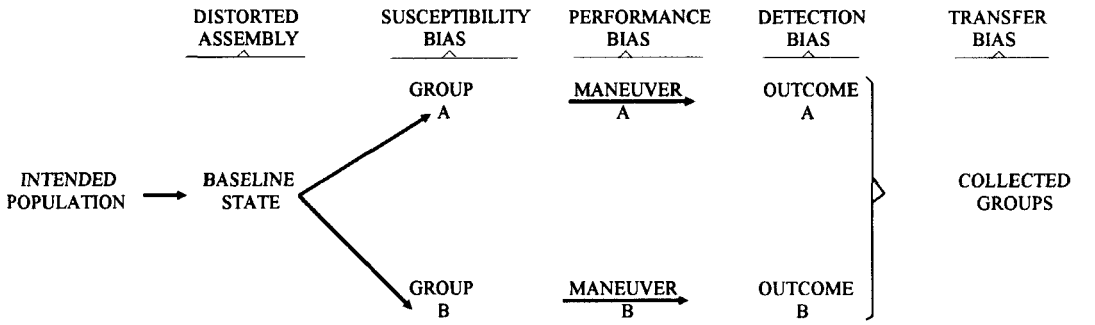

Figure 1. Feinstein's 'intellectual model' to evaluate the scientific quality of cause–effect research

hold over populations in which averaging has occurred and in which homogeneity is neither guaranteed nor even strongly warranted. This is a different kind of claim than causal claims that are grounded on entity identity or universality. In those situations causal generalizations apply to each and every individual at the level of the individual, whereas a population-level causal claim has validity only for groups over which variations have been averaged.

This view does not preclude us, since we may adopt an analogue of the Reichenbach and Salmon[44] notion of 'weight', from fictitiously or 'hypothetically' ascribing a kind of putative individual causation based on population causation through the probabilistic concept of odds or risk. This tack allows some degree of manipulability and permits predictions to be made and treatment regimens designed, though these are actuarial types of predictions, and are strictly speaking empirically applicable to averaged populations. Because of the type of prediction and control that such actuarial generalizations permit, and because the generalizations may well associate properties temporally, that is, E follows temporally after C, such generalizations exhibit important analogies to fundamental or basic causal generalizations. It is important to keep in mind, however, that the forms of putative individual causation, population causation, and basic causation (that is, the type meeting Cartwright's CC condition) are only analogous, and that they have their domain of application at two different levels, hold in different populations, and are based on different test conditions. These varying *concepts* of causation do not, in my view, require different underlying metaphysical accounts of causation: one will suffice even though there may be different ways that such causation is tested for, captured, and represented. See Giere's[28] account as presented above and also my discussions in my recent essay[33] and book.[12] (The only situation in which the metaphysical underpinnings of our causation concepts may possibly require a new foundation is, I think, in quantum mechanical causation, where there is evidence that nature presents us with 'stochastic bedrock'. That excepted, the account of Giere discussed in the text should provide the basis for a deterministic sense of causation consistent with Mackie's analysis[20] that also permits 'probabilistic causation'.)

A somewhat similar set of considerations has been urged by several statisticians who have considered causation and clinical trials, including Rubin,[4,5] Holland[8] and Cox.[45] I will discuss these views and how they relate to the position developed in the present section after I have introduced Rubin's analysis in the following section.

6. DRAWING THE THREADS TOGETHER: A BAYESIAN APPROACH

Immediately after Cartwright proposes her CC criterion, she writes that 'the practical difficulties with this criterion are conspicuous.'[1] A bit further on in her book she elaborates:

> the method does not literally require one to know all the other causes. Rather what you must know are some facts about what the other probabilities are in populations that are homogeneous with respect to all these other causes, and that you can sometimes find out without first having to know what the other causes are. This is the point of the randomized experiment.[1]

A randomized experiment for Cartwright is however a different creature than those who conduct clinical trials implement, for an actual clinical trial is at best an approximation to *an ideal experiment*. As I read Cartwright, actual experiments generate frequencies, but these – *contra* Giere[28] as discussed above – are not sufficient to license probabilities without the fulfilment of additional conditions. She writes: 'It is not frequencies that yield causes, but probabilities; and it is not results in real experiments, where subjects are assigned to groups by a table of random numbers, but rather in ideal experiments where randomization is actually achieved.'[1]

I think the point here is based on a belief (better a hope) that true randomization will control for *any* bias not controllable on the basis of known information. Interestingly Glymour and his colleagues' recent development of their approach to causal modelling also views randomization as an important tool to control faulty causal inferences, writing that 'Random assignment of units to experimental treatment categories ensures there is no such [confounding common] causal connection; randomization guarantees that in the sample created by experiment other causes of Y are not also causes of X' (where the experiment seeks to determine the effect of a treatment X on an effect Y). But this guarantee as we have heard from some Bayesians is a suspect thesis,[46] and Glymour elsewhere admits that sample variation in any actual sample may void this guarantee, which might only be approached asymptotically through long-term reiterated studies (personal communication). I want to close this paper by referring, however, to one Bayesian, Donald Rubin, who has defended the role of randomization, since I think that his position comports well with both the position advanced by Cartwright as well as the theses developed in this paper. Importantly Rubin's framework has also very recently been unified with the causal modelling approach by Glymour and his associates.

Rubin's account proposes that we think of a study of T treatments as a complex matrix designed to represent in principle *all* potentially observed values (see Figure 2). For Rubin, 'causal effects are comparisons among values that would have been observed under all possible assignments of treatments to experimental units.'[4] In addition, the matrix will contain information about the pre-treatment values, which treatment the patient received (W), post-treatment values, and indicators for any missing data (M). In controlled experiments, the 'which treatment' column reflects *two* mechanisms: the sampling mechanism (who gets studied) and the treatment assignment mechanism. For Rubin the determination of causation is predicated on having *all possible* values observed that, if available, would allow the *calculation* of causal effects. Since this is, he adds, 'impossible', we must instead employ (Bayesian) statistical inference to *estimate* causal effects.

This statistical inference will, however, only work if the assignment mechanism (W) and the recording mechanism (M) are 'ignorable'. On Rubin's analysis, for all but very simple (and artificial) situations, that is in practical situations, a solution is to employ randomization for the W data. This, Rubin argues, can 'markedly reduce the sensitivity of a valid Bayesian analysis, because only a randomized assignment mechanism can be ignorable and yield data having more than one treatment condition represented for a distinct value of recorded covariates'.[4] Using an alternative deterministic assignment rule would lead, he adds, to difficulties in both study execution and analysis.

Rubin's[4] analysis discussed above has been extended in further work by him[5] and also in a joint address with Holland[47] and by Holland himself.[8] Of special relevance to the thesis

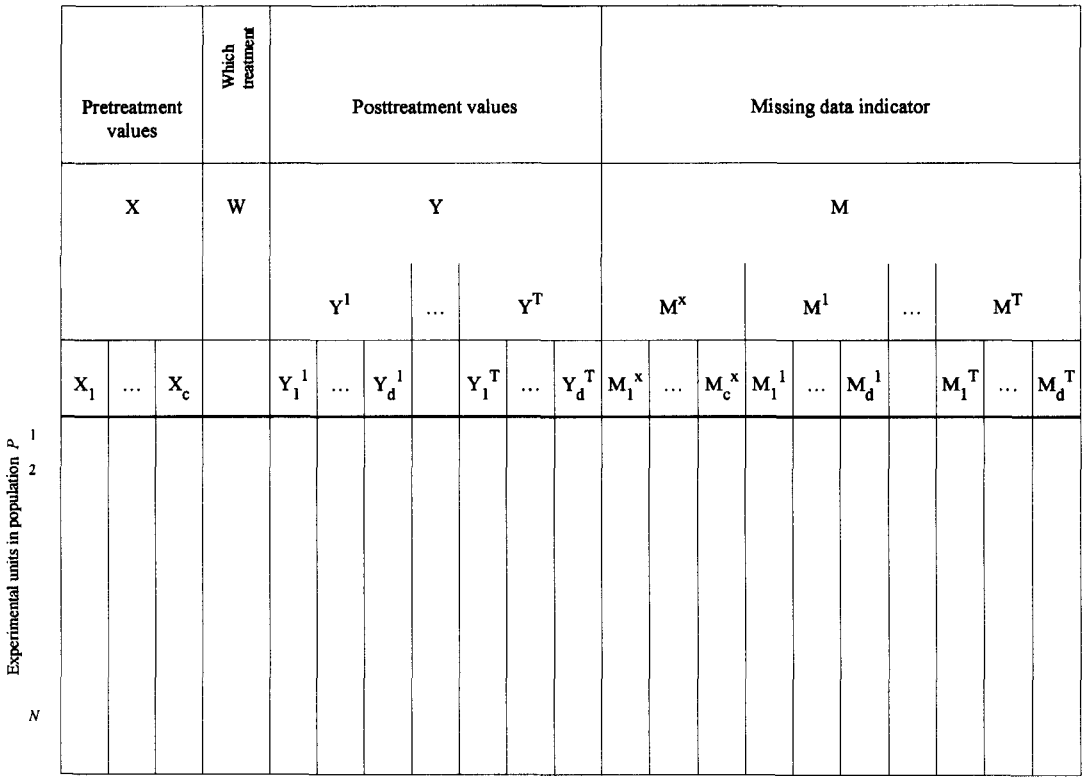

Figure 2. All values in a study of T treatments

developed in the previous section concerning two concepts of causation, the physiological and the epidemiological, are Holland's amplifications of Rubin's views concerning the homogeneity of the entities investigated in laboratory science versus clinical trials. Holland introduces what he terms the *fundamental problem of causal inference* as follows:

> It is impossible to *observe* the value of [the response variable for the treated unit] $Y_t(u)$ and [the response variable for the control unit] $Y_c(u)$ on the same unit and, therefore, it is impossible to *observe* the effect of t on u.[8]

If the units are identical, however, it is possible to make the causal inference and conclude that treatment t causes the effect represented by the difference $Y_t(u) - Y_c(u)$; this is what Holland terms the 'scientific' solution. (The conditions under which one would find such unit identity are discussed in one of my recent essays.[33]) When this condition of identity or homogeneity of units is *not* satisfied, one may, following Holland's explication of Rubin, still draw a causal conclusion, but now of a *statistical* sort. Holland writes:

> The statistical solution is different and makes use of the population U in a typically statistical way. The *average causal effect*, T, of t (relative to c) over U is the expected value of the difference $Y_t(u) - Y_c(u)$ over the u's in U; that is,
>
> $$E(Y_t - Y_c) = T. \tag{3}$$

T defined in (3) is the average causal effect. By the usual rules of probability (3) may also be expressed as:

$$T = E(Y_t) - E(Y_c). \tag{4}$$

Although this does not look like much, (4) reveals that information on *different* units that can be *observed* can be used to gain knowledge about T.[8]

Holland elaborates somewhat further, contrasting the *average* causal effect of T on the units in U with the causal effect of T on a specific unity u_0. He then introduces the assumption of *constant effect* which, though weaker than the assumption of unit homogeneity, allows the inference that

$$T = Y_t(u) - Y_c(u), \qquad \text{for all } u \text{ in } U.$$

This assumption is also known as the assumption of *additivity* since it expresses the claim that treatment t adds a *constant* amount T to the control response for *each and every* unit. This additivity assumption is not new, but rather has a venerable history in the literature on clinical trials. Holland also notes that this additivity assumption can be tested by dividing U into subpopulations: if the Ts vary in the subpopulations then the assumption does not hold. Holland views the constant effect or additivity assumption as a weakening of the assumption of unit homogeneity, but in my view *if* we confine our attention to the factors of interest (the Ys) and perform a complete subgroup analysis, which will be tantamount to examining each individual unit, the distinction between the two assumptions essentially collapses. The additivity assumption is an important one, however, and if interpreted as applying to the *specific* factors of interest in an epidemiological investigation (including randomized controlled clinical trials) does not require *complete* homogeneity of each of the units being compared. It is presumably this feature that led Sir David Cox to stress the importance of this assumption, which he terms 'unit-treatment additivity' in his comments[4,5] on Holland's article.

I interpret Holland's account of the difference between the scientific solution and the statistical solution to the fundamental problem of causal inference, as well as his approach to the assumption of unit-treatment additivity, as focusing on the same type of distinction I drew in the previous section where I contrasted physiological with epidemiological approaches to causation.

Rubin's framework, which Holland elaborates in his remarks cited above, also seems compatible with Cartwright's position. Interestingly Rubin's approach has in addition been very recently unified with the causal modelling approach by Glymour and his colleagues.[3] I will not have an opportunity in the present paper to describe the causal modelling approach introduced and developed by Kiiveri and Speed,[48] Wermuth,[49] Lauritzen and Wermuth,[50] Pearl[51] and Glymour *et al.*,[3] but references to this literature are provided to enable the interested reader to access this domain. Glymour has recently generalized the model developed in his and his associates' 1987 volume on *Discovering Causal Structure*[2] and now has founded that approach on four conditions they term the Markov, Minimality, Faithfulness, and Manipulation conditions. Though space does not permit a statement of these conditions in this paper, suffice it to note that Glymour and his colleagues claim:

> Rubin's analysis of treatment assignment determined by a covariate gives exactly the result that would be obtained by applying the Manipulation condition. That seems to us good reason to think that the structure the Rubin framework is after is caught by the Markov, Faithfulness and Manipulation conditions.[3]

It seems to me that the arguments developed by Glymour and his colleagues are sound ones, and that their unification offers much promise for probing the foundations of what I term clinical

causation. In addition, the unification may point the way towards resolution of other disagreements about the foundations of the relations of probabilities and causes between Glymour and Cartwright as well.

It is not the task of the present paper to investigate the issues of experimental design or of randomization in depth; that role fell to Peter Urbach and others who address these matters in their papers. From the perspective of this paper it seemed useful to show the similarities contained in the CC criterion to Rubin's Bayesian approach to causation, to indicate that at least one Bayesian found that a practical solution to the innumerable causes/cases problem was also achievable through randomization, and that recent unifications of Rubin's framework by Glymour and his colleagues point the way to additional promising foundational work on these issues. As I interpret it, randomization is insurance against bias, and will give us our best shot at blocking spurious causal attributions, given that we have adequately controlled for known confounders, something that Imre Lakatos might have said would be in accord with his belief that there was no *instant* rationality in science.[52] This, I think, is ultimately the message of the epidemiologists, and it is one which I see as consistent with the deepest recent philosophical work on causation as well.

7. SUMMARY AND CONCLUSION

In this article I have argued that a manipulability approach to causation is the fundamental sense of causation and that this type of analysis is committed to an appeal to counterfactual considerations. Further, I tried to show that the type of causal claims that epidemiologists *and clinical trialists* make do not satisfy the ideal conditions, represented by Cartwright's CC principle, though both scientific or physiological causation as well as probabilistic causation discussed by philosophers do meet this condition. I characterized the kind of causation licensed by epidemiological studies including well-designed clinical trials as holding at the level of the population, and only inaccessibly at the level of the individual. How this is possible was illustrated by using Giere's analysis of deterministic individual systems which exhibit probabilistic causal relations at the population level. I argued that physiological causation could be either deterministic or stochastic, illustrating both forms with examples from neuromuscular physiology, and maintaining that these systems could meet the CC condition. I reviewed attempts to warrant causal claims in epidemiology including clinical trials, identifying criteriological and design-based approaches, and suggesting that the latter represented an endeavour at achieving the type of homogeneity required by the CC principle, but not one that could be expected to be attained. Finally I embedded this analysis in a Bayesian approach to causation pioneered by Rubin and developed further by Holland, and linked this analysis to recent causal modelling accounts of Glymour and his associates. The extent to which this Bayesian framework can be still further developed and joined with various causal modelling approaches to improve our analyses of scientific and epidemiological causation points to important future work at the intersection of biology, medicine, and statistics.

REFERENCES

1. Cartwright, N. *Nature's Capacities and their Measurement*, Oxford University Press, New York, 1989.
2. Glymour, C., Scheines, R., Spirtes, P. and Kelly, K. *Discovering Causal Structure*, Academic Press, San Diego, 1987.
3. Spirtes, P., Glymour, C. and Scheines, R. *Causation, Prediction and Search*, Lecture Notes in Statistics, Springer, 1992.
4. Rubin, D. B. 'Bayesian inference for causal effects: the role of randomizations', *Annals of Statistics*, **6**(1), 34–58 (1978).

5. Rubin, D. B. 'Comment: which ifs have causal answers', *Journal of the American Statistical Association*, **81**, 961–962 (1986).
6. Pratt, J. W. and Schlaifer, R. 'On the nature and discovery of structure', *Journal of the American Statistical Association*, **79**, 9–33 (1984).
7. Pratt, J. W. and Schlaifer, R. 'On the interpretation and observation of laws', *Journal of Econometrics*, **39**, 23–52 (1988).
8. Holland, P. W. 'Statistics and causal influence', *Journal of the American Statistical Association*, **81**(396), 945–960 (1986).
9. Lewis, D. 'Causation', *Journal of Philosophy*, **70**, 556–557 (1973).
10. Brand, M. (Ed.) *The Nature of Causation*, University of Illinois Press, Urbana, IL, 1976.
11. Earman, J. *A Primer on Determinism*, Kluwer, Dordrecht, 1986.
12. Schaffner, K. F. *Discovery and Explanation in Biology and Medicine*, University of Chicago Press, Chicago, 1993.
13. Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. *Statistical Package for the Social Sciences*, 2nd edition, McGraw-Hill, New York, 1975.
14. Hart, H. L. A. and Honoré, A. M. *Causation in the Law*, 2nd edn, Oxford University Press, Oxford, 1985 (1959).
15. Collingwood, R. G. *An Essay on Metaphysics*, Clarendon Press, Oxford, 1940.
16. Gasking, D. 'Causation and recipes', *Mind*, **6**, 479–487 (1955).
17. von Wright, G. H. *Explanation and Understanding*, Cornell University Press, Ithaca, NY, 1971.
18. Glymour, C. 'Comment: statistics and metaphysics', *Journal of the American Statistical Association*, **81**, 964–966 (1986).
19. Howson, C. and Urbach, P. *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle, IL, 1989.
20. Mackie, J. *The Cement of the Universe*, Oxford University Press, Oxford, 1974.
21. MacMahon, B. and Pugh, T. F. *Epidemiology: Principles and Methods*, Little Brown, Boston, 1970.
22. Fletcher, R. H., Fletcher, S. W. and Wagner, E. H. *Clinical Epidemiology – the Essentials*, Williams and Wilkins, Baltimore, 1982.
23. Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. (1982) *Epidemiologic Research*, Wadsworth, London, 1982.
24. Reichenbach, H. *The Direction of Time*, University of California Press, Berkeley, 1956.
25. Good, I. J. 'A causal calculus I–II', *British Journal for the Philosophy of Science*, **44**, 305–318 and **45**, 43–51; 'Errata and corrigenda', **49**, 88 (1961–62).
26. Suppes, P. *A Probabilistic Theory of Causality*, North-Holland, Amsterdam, 1970.
27. Salmon, W. 'Probabilistic causality', *Pacific Philosophical Quarterly*, **61**, 50–74 (1980).
28. Giere, R. 'Causal systems and statistical hypotheses', in Cohen, L. J. and Hesse, M. B. (eds.), *Applications of Inductive Logic*, Oxford University Press, Oxford, 1980, pp. 251–270.
29. Cartwright, N. 'Causal laws and effective strategies', *Nous*, **13**, 419–437 (1979).
30. Cartwright, N. *How the Laws of Physics Lie*, Oxford University Press, New York, 1983.
31. Sober, E. *The Nature of Selection*, MIT Press, Cambridge, MA, 1984.
32. Eells, E. and Sober, E. 'Probabilistic causality and the question of transitivity', *Philosophy of Science*, **50**, 35–57 (1983).
33. Schaffner, K. F. 'Causing harm: epidemiological and physiological concepts of causation', in Mayo, D. G. and Hollander, R. (eds.), *Acceptable Evidence: Science and Values in Hazard Management*, Oxford University Press, New York, 1991.
34. Schaffner, K. F. 'Philosophy of method', in Lederberg, J. (ed.) *Encyclopedia of Microbiology*, Academic Press, San Diego, 1992.
35. Schaffner, K. F. 'Theory structure in the biomedical sciences', *The Journal of Medicine and Philosophy*, **5**, 57–97 (1980).
36. Schaffner, K. F. 'Exemplar reasoning about biological models and diseases: a relation between the philosophy of medicine and philosophy of science', *Journal of Medicine and Philosophy*, **11**, 63–80 (1986).
37. Katz, A. M. *Physiology of the Heart*, Raven Press, New York, 1977.
38. Kuffler, S. W., Nicholls, J. and Martin, A. *From Neuron to Brain*, 2nd edn, Sinauer Associates, Sunderland, MA, 1984.
39. Wimsatt, W. 'Reductionism, levels of organization, and the mind–body problem', in Globus, G., Maxwell, G. and Savodnik, I. (eds.), *Consciousness and the Brain*, Plenum Press, New York, 1976, pp. 205–267.

40. Hill, Sir A. B. *Principles of Medical Statistics*, 8th edn, Oxford University Press, Oxford, 1966.
41. Evans, A. S. 'Causation and disease: the Henle–Koch postulates revisited', *Yale Journal of Biology and Medicine*, **49**, 175–195 (1976).
42. Lilienfeld, A. M. and Lilienfeld, D. E. *Foundations of Epidemiology*, 2nd edn, Oxford University Press, New York, 1980.
43. Feinstein, A. R. *Clinical Epidemiology: The Architecture of Clinical Research*, Saunders, Philadelphia, 1986.
44. Salmon, W. *The Foundations of Scientific Inference*, University of Pittsburgh Press, Pittsburgh, 1967.
45. Cox, Sir D. R. 'Comment', *Journal of the American Statistical Association*, **81**, 963–964 (1986).
46. Urbach, P. 'The value of randomization and control in clinical trials', *Statistics in Medicine*, **12**, 1421–1431 (1993).
47. Holland, P. W. and Rubin, D. B. 'Causal inference in prospective and retrospective studies', address given at the Jerome Cornfield Memorial Session of the American Statistical Association Annual Meeting, August 1980.
48. Kiiveri, H. and Speed, T. 'Structural analysis of multivariate data: a review', in Leinhardt, S. (ed.), *Sociological Methodology*, Hoeesy Bass, San Francisco, 1982.
49. Wermuth, N. 'Linear recursive equations, covariance selection and path analysis', *Journal of the American Statistical Association*, **75**, 963–972 (1980).
50. Lauritzen, S. and Wermuth, N. 'Graphical models for associations between variables, some of which are qualitative and some quantitative', *Annals of Statistics*, **17**, 31–57 (1989).
51. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*, Morgan and Kaufman, San Mateo, CA, 1988.
52. Lakatos, I. 'Falsification and the methodology of scientific research programmes', in Lakatos, I. and Musgrave, A. (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, 1970, pp. 91–196.