SOBER’S PRINCIPLE OF COMMON CAUSE AND THE PROBLEM OF COMPARING INCOMPLETE HYPOTHESES*

MALCOLM R. FORSTER†

Department of Philosophy
University of Wisconsin–Madison

Sober (1984) has considered the problem of determining the evidential support, in terms of likelihood, for a hypothesis that is incomplete in the sense of not providing a unique probability function over the event space in its domain. Causal hypotheses are typically like this because they do not specify the probability of their initial conditions. Sober’s (1984) solution to this problem does not work, as will be shown by examining his own biological examples of common cause explanation. The proposed solution will lead to the conclusion, contra Sober, that common cause hypotheses explain statistical correlations and not matchings between event tokens.

1. Introduction. Van Fraassen (1980, 1982) has argued that the Reichenbach/Salmon schema for common cause explanation accords poorly with standard theoretical treatments of quantum mechanical phenomena. In response, Sober (1984) has applied likelihood methods in order to understand better the rationale behind common cause explanation. (For our purposes it is sufficient to define likelihood of a theory as the probability of the data given the theory.) The problem is that theories commonly do not give a definite probability to the events in their domain of explanation.

This problem appears to be relevant to the scientific realism debate. One prominent argument for realism (Smart 1963, 1985) appeals to the idea that observed regularities would remain nothing more than a “cosmic coincidence” were we not to believe in a reality behind the phenomena. In other words, the best explanation of the phenomena should be believed because it makes the observed data more probable than they would be otherwise. But it is a nontrivial question to ask whether the “best explanation” can be shown to have higher likelihood than the “null” hypothesis.

Applying likelihood methods to causal explanations is not straightfor-

*Received June 1986; revised September 1986.
†I am very grateful to Elliott Sober for extensive correspondence on these issues. Some crucial mistakes were corrected in an earlier draft, thanks to Nancy Cartwright. A substantial part of this work was carried out under a University Post-Doctoral Fellowship at Monash University, Australia.

Copyright © 1988 by the Philosophy of Science Association.

538
ward, because such hypotheses do not make exact predictions about the probabilities of observed events, or collections of events. For example, it is obvious that while Newtonian mechanics may say something about the probability of the final state of a ball rolled down a slope given the initial state, it usually says nothing about the probability of the initial (or final) stages unconditionally. Common cause explanations are incomplete as well—partly because they do not specify the probability of the common cause events, but perhaps also because the transition probabilities from past to future events are not precisely known. So, we typically do not have an exact numerical value for the probability of the observations (data set) given the theory, and so we don’t have a likelihood value to measure evidential support either.

Sober (1984) provides one method of comparing “incomplete” theories against the evidence. He then proceeds to use his idea to formulate a replacement for the Reichenbach/Salmon principle of common cause, and he argues that his formulation does not founder upon the rocks of quantum mechanics. Unfortunately, Sober’s new principle of common cause is not successful. The fault lies with a small part of his framework, namely, the strategy he chooses to make comparisons between probabilistically incomplete theories. Because of this difficulty, his solution to the problem of comparing incomplete theories does not work, as will be argued in section 4. My own solution to the problem appears in section 5, and is applied to Sober’s biological examples in section 6.

The basic idea behind my proposal is easy enough to explain. Probabilistically incomplete theories can be represented as (non-singleton) sets of probability functions. We should define the likelihood of the theory as the likelihood of the element with maximum likelihood, and then we should choose that probability function as representing the theory. An immediate consequence of this approach is that Sober is wrong when he says that common causes explain matchings between token events (section 3). Rather, theoretical entities tend to explain more global statistical features of data collected from a variety of different situations, and in the case of common cause the statistic explained is the correlation between two variables (section 6). This conclusion fits well with the intuitive examples presented by Sober himself, as will be pointed out in my concluding remarks (section 7).

2. Common Cause Explanation. Roughly speaking, Reichenbach’s (1956) principle states that any statistically significant correlation observed between two variables pertaining to simultaneous (space-like separated) events should be explained by postulating the existence of a common cause variable referring to events in their shared past (the intersection of their past lightcones). (Common causes may be introduced with respect
to non-simultaneous events, but there is no imperative in this case.) For example, what appears on the screen of one television set is correlated to what appears on another set tuned to the same channel. We don’t explain this in terms of any mysterious causal connection between them; we postulate a mechanism by which both television sets respond to a common electromagnetic signal somewhere in their past. The precise formulation of Reichenbach’s principle is (van Fraassen 1980, p. 28—I have substituted “.” for “&”):

If coincidences of two events A and B occur more frequently than would correspond to their independent occurrence, that is, if the events satisfy the relation

$$P(A.B) > P(A).P(B),$$

then there exists a common cause C for these events such that the fork ACB is conjunctive, that is satisfies relations (2)–(5) below:

$$\begin{align*}
(2) & \quad P(A.B/C) = P(A/C).P(B/C) \\
(3) & \quad P(A.B/\bar{C}) = P(A/\bar{C}).P(B/\bar{C}) \\
(4) & \quad P(A/C) > P(A/\bar{C}) \\
(5) & \quad P(B/C) > P(B/\bar{C})
\end{align*}$$

The premises (2)–(5) then explain the fact of (1) in that (1) follows logically from those premises.

It is generally agreed that not all correlations in science are explained in this way, and nor should they be. As van Fraassen (1980) has been keen to point out, some quantum mechanical correlations cannot be explained by the introduction of common causes (hidden variables) as Reichenbach’s principle demands. Given that local hidden variables are incompatible with quantum mechanics, as Bell (1964) argued, Reichenbach’s principle of common cause does not apply to this example (van Fraassen 1982). Even Salmon’s (1980) generalization of Reichenbach’s principle in terms of interactive forks\(^1\) is badly damaged as a reply to this criticism by the fact that even non-local hidden variables (though compatible with quantum mechanics) are not generally insisted upon by the scientific community at large. It follows that Salmon’s generalization accords poorly with actual scientific practice in these cases.

Examples from classical physics also commonly fail to conform to the Salmon/Reichenbach schema (Forster 1986a). Sober (1984) has also put forward some examples of phylogenetic inference in biology that allegedly violate the Reichenbach/Salmon common cause principles (but I will dispute that later).

\(^1\)Salmon relaxes the conditional independence assumptions (2) and (3). In effect, this allows a direct (non-local) causal interaction between the end of the fork—that is, between A and B—hence the name “interactive” fork.
So, the important issue raised is whether we can find another principle of common cause that dictates, in a non ad hoc way, the limits of its own application. Sober (1984) seems to have put forward a proposal along these lines. He first considers examples of phylogenetic inference concerning the ancestry of a group of sister species based on the observed similarities and differences between them. These examples have to be studied before Sober's proposed principle of common cause can be properly understood.

However, my basic source of disagreement with Sober can be introduced now. Many authors say that common cause hypotheses explain correlations, but it is not always clear what they mean by 'correlation'. That word is commonly used in two crucially different senses, as we shall see in the next section. Sober uses the word to refer to "matchings of token events", but I will argue that these are not what are explained by common cause hypotheses at all. Rather, I claim that common causes explain correlations in the statistical sense of the word.

3. Two Meanings of 'Correlation'. In its more technical statistical usage the term 'correlation' refers to a measure of association between two (or more) variables. For example, if we take one variable to record some property of the pattern appearing on one screen and a second variable to record the same property on the second screen, then if there is a high degree of association when we compare values on a number of different occasions we talk of a statistical correlation between the variables: one variable is a function of the other, to use the more familiar mathematical jargon. Whenever we have a correlation between two variables, knowing the value of one variable is useful in predicting the value of the other. Consider now a joint probability measure \( P(A = i, B = j) \), where \( i \) and \( j \) range over all possible values of the variables (for simplicity we are taking the variables to be discrete, but generalization to the continuous case is straightforward). The technical measure of (linear) correlation, rho, is defined as:

\[
\rho(A,B) = \frac{d}{f} \frac{\text{Cov}(A,B) / \sigma(A) \cdot \sigma(B)}{, (1)}
\]

where \( \sigma(A) = (\text{Cov}(A,A))^{1/2} \), \( \sigma(B) = (\text{Cov}(B,B))^{1/2} \). The covariance is defined as:

\[\text{Cov}(A,B) = \text{E}[(A - \text{E}(A))(B - \text{E}(B))]\]

It is important to note that a correlation in this sense is a relationship between variables and not between events. This fact is not obvious from considering the special case of dichotomous variables, and this has been the source of some confusion in the past. Thinking of a correlation as a functional connection helps avoid this mistake.
\[
\text{Cov}(A,B) = \sum_i \sum_j P(A = i, B = j)i \cdot j - \left[ \sum_i P(A = i) \right] \left[ \sum_j P(B = j) \right].
\]  

(2)

The variables A and B can be thought of as vectors in an inner product space, under certain restrictions, with inner product Cov(A,B) and magnitudes \(\sigma(A)\) and \(\sigma(B)\) respectively. The correlation coefficient \(\rho(A,B)\) is then the cosine of the angle between those vectors, as is seen to be plausible by the fact that \(-1 \leq \rho(A,B) \leq 1\,\text{, and }\rho(A,B) = 0\,\text{ if }A\text{ and }B\text{ are statistically independent (that is, orthogonal). (See Eaton 1983, for a rigorous mathematical treatment).}

In the simple case in which two variables, A and B, take on only two possible values, say 0 and 1, the covariance of the two-valued variables (often called dichotomous variables) reduces to:

\[
\text{Cov}(A,B) = P(A = 1, B = 1) - P(A = 1) \cdot P(B = 1).
\]  

(3)

When we speak of observed correlations the probability function \(P\) refers, then, to observed relative frequencies, which we will denote by \(r(A = i, B = j)\). The covariance, Cov(A,B), is often itself used as a measure of association because it always has the same sign as \(\rho(A,B)\). That is, Cov(A,B) \(\geq 0\) iff \(\rho(A,B) \geq 0\), Cov(A,B) = 0 iff \(\rho(A,B) = 0\) (provided that \(\sigma(A) \neq 0\) and \(\sigma(B) \neq 0\)), and Cov(A,B) \(\leq 0\) iff \(\rho(A,B) \leq 0\). Other measures that have the same sign as \(\rho(A,B)\) are \(P(A = 1/B = 1) - P(A = 1)\) and \(P(B = 1/A = 1) - P(B = 1)\), as well as the regression coefficients \(P(A = 1/B = 1) - P(A = 1/B = 0)\) and \(P(B = 1/A = 1) - P(B = 1/A = 0)\). (We arrive at the same notation as used by Reichenbach and van Fraassen—quoted earlier—if we write \(A\) as \(A\) and \(A = 0\) as \(\tilde{A}\), etc. But my notation has the advantage of not conflating the variable \(A\) with the event \(A\). Because this distinction is crucial to the subject of this paper, I will use the longer notation.)

The other sense of ‘correlation’, which might be called event correlation, is given by the layman’s use of the word to denote a matching between two singular event tokens. So, if we observe the simultaneous events \(A = 1\) and \(B = 1\), we say that these two events are correlated. If we observe many joint occurrences of variables \(A\) and \(B\), then the degree of event correlation is naturally measured by the relative frequency \(r(A = 1, B = 1)\). If only one instance is observed then the event correlation is either 1 or 0.

The two senses of ‘correlation’ are not unrelated. If many instances are observed then a high statistical correlation implies that joint occurrences \((A = 1, B = 1)\) are not uncommon. But the converse definitely
does not hold. Also, if only one instance of A and B having the value 1 is observed then the two concepts diverge since event correlation is high, but the statistical covariance is zero and the statistical correlation is undefined. In less trivial cases, it is still possible that \( p(A, B) \) is low, even zero, when \( r(A = 1, B = 1) \) is high, even 1. Imagine drawing a marble from each of two urns (with replacement) each 90% filled with black marbles and 10% with white marbles. Then, in the long run, over 80% of the draws will produce matchings of two black marbles, while statistical correlation will be zero. \( r(A = 1, B = 1) \) is high but so is \( r(A = 1), r(B = 1) \), so their difference—the covariance \( \text{Cov}(A, B) \)—is close to zero. The quantity \( r(A = 1), r(B = 1) \) is the proportion of joint occurrences of \( A = 1 \) and \( B = 1 \) we would expect on the basis of independence—that is, on the basis of assuming that each draw is independent of the other. It is only when the actual proportion of joint occurrences \( r(A = 1, B = 1) \) exceeds this quantity \( r(A = 1), r(B = 1) \) that we have a positive statistical correlation.

Sober distinguishes these two meanings of ‘correlation’ and picks the “matching of single token events” as the sense of the word appropriate to his principle of common cause:

Two possible explananda must be distinguished. One may set out to explain a known correlation between kinds of events in a given population; or one may wish to explain the matching of single token events in that population. The literature on the principle of common cause frequently treats these two projects interchangeably. But, as noted above, a matching of token events may be observed even when no correlation of types is known to exist. The principle of common cause will be construed in what follows to apply to this sort of situation. (Sober 1984, p. 220)

Admittedly, there are two ways of understanding what Sober might mean by a ‘correlation between kinds of events in a given population’. He may mean a statistical correlation measured by \( r(A = 1, B = 1) - r(A = 1), r(B = 1) \), or the relative frequency of matching in the population given by \( r(A = 1, B = 1) \). If he means the latter then I will argue that both of his alternatives are wrong. If he means the former then he made the wrong choice between them. In my view, common cause explanation correctly applies to situations in which a statistical correlation holds between two variables.² It is a mistake to apply any principle of common

²Certainly, Reichenbach (1956) would disagree with Sober’s choice, although he appears not to pay much attention to the distinction at all. But it is clear from the technical details of his explanatory schema that a common cause is meant to explain a statistical covariance. For, the explanandum of his schema is stated as \( [P(A, B) - P(A)P(B)] > 0 \), and not as \( A , B \) or \( P(A, B) > 0 \).
cause to a situation in which no statistical correlation is known to exist. For in such a situation we have no evidence about whether the variables, which are instantiated by these token events, are functionally connected or not. And if they are not, then there is no need to postulate any common cause to explain their token co-occurrence.

If I toss two coins and they both come up heads, there is no justification for supposing that these two events have a common cause. If I draw black marbles from two different urns, there is no need to explain this co-occurrence in terms of a common cause. This point was made by Aristotle in *Physics*, Book II, chapter 5, and cited by van Fraassen (1980, p. 25) and Cartwright (1983, p. 201). Aristotle imagines a chance meeting of two men at the market. The meeting is fortuitous because each man goes to the market for independent reasons. The meeting is still a physical fact, and it has an explanation. But its explanation is complete when we have finished explaining each man’s trip to the market separately. There is nothing more to be said about the *meeting* that is not already said in explaining why the first man was there and why the second man was there.

But if each man went to the market at irregular times just once a week and kept on meeting week after week, then we would suspect that there is something more to be said. Perhaps both men have a common incentive in choosing when they go to market—perhaps they are both fond of the fishmonger’s daughter, who turns up once a week to help unload the catch. It does not suffice to explain their *continual* meetings to state that each man goes to the market to see a girl—we also need to mention that they go to see the *same* girl. We then have a typical case of common cause explanation.

The point being argued is that there is nothing significant in matchings between events per se. In most explanatory contexts, matchings between events do not need explaining in terms of common causes. Another way of seeing this is to notice that, in a sense, matchings *always* occur—they are too common to be significant in themselves. Take the example of drawing marbles from two urns, and suppose we draw a black marble with a white one. If we define the *anti-color* of a marble to be the color that it’s not (that is, black if it’s white and white if it’s black), then the *anti-color* of the first marble matches the *color* of the second marble on this occasion. But, according to Sober’s own principles, he should take this as good evidence for the existence of a common cause.

Sober might want to get around this counterexample by denying that anti-color is a genuine property, and so claim that there is no matching between *genuine* properties in this case, and so no need to (wrongly) posit a common cause. But then he cannot account for clear-cut examples of common cause inference in which no “genuine” matchings occur. For
consider the case in which we observe, over time, the colors of two sets of traffic lights at an intersection. We notice, of course, that one set is red whenever the other set is green and vice versa. Although we correctly take this as excellent evidence for the existence of a common cause, there are no “genuine” color matchings observed. Whence, we prove that “genuine” matchings between events are neither a sufficient condition for rational common cause inference (as shown by the urn example), nor a necessary condition (as shown by the traffic lights example).

To be fair to Sober, his position is more complicated than indicated so far. He rejects the question “When do pairs of events have common causes?” because he believes that it has a trivial answer: “always” (Sober 1984, p.220). So, he considers three event-types A, B, and C, and asks when one pair of events has a more recent common cause than the others. Secondly, he adds other theoretical assumptions in his examples common to all competing hypotheses, and this may make a difference. Thirdly, he evokes a certain strategy to compare hypotheses with incomplete parameter specifications (in particular, unknown transition probabilities), and this needs to be critically examined. So, an adequate critique of Sober’s principle of common cause must wait until a later section. But in the meantime the reader should have a good indication of its basic fault.

4. Sober on Phylogenetic Inference. Sober (1984) chooses biological examples in which three events A = i, B = j, and C = k, are assumed to share a common cause sometime in the past. The question is then whether or not the events A = i and B = j have a more recent common cause than that shared by all three events. Hypothesis (AB)C says “yes”, while (ABC) answers “no”.

In Figure 1, (AB)C is depicted as a common cause (CC) explanation of any matching between the events A = i and B = j, while (ABC) is shown as the surrogate for a single cause (SC) explanation of that matching. (AB)C posits a common cause variable D (whose exact value is left unspecified) while (ABC) does not. Given certain assumptions, Sober shows
that we can discriminate between \((AB)C\) and \((ABC)\) on the basis of likelihood. (For our purposes, the *likelihood* of a hypothesis is defined to be an unspecified constant \(k\) times the probability of the data given that the hypothesis is true, where \(k\) depends on the data only. Likelihood *comparisons* among hypotheses can be made, but only relative to the same data set—see Edwards 1972.)

In his examples, Sober compares three biological populations \(A, B,\) and \(C\) with respect to one characteristic which can be in one of two possible character states—primitive (denoted by 0) or advanced (denoted by 1). That is, we consider three variables \(A, B,\) and \(C,\) representing the character states (of the one characteristic) of the populations \(A, B,\) and \(C,\) respectively. These variables take one of the values in the set \(\{0,1\}.\) The hypotheses \((AB)C, A(BC),\) and \((ABC)\) tell us *something* about the causes of present character states, and thereby *something* about the past phylogenetic history of the three populations. The question at hand is "How do we compare these hypotheses against the evidence?" This cannot be answered until more is said about what these hypotheses entail. In particular, we want each hypothesis to place some constraints on the probabilities of joint occurrences of events, such as \((A = i.B = j.C = k)\) (where the "\(\cdot\)" stands for "and"). This is done indirectly by constraining the theoretical *transition* probabilities such as \(P(A = 1/D = 1)\) or \(P(B = 1/E = 0).\) For the example of phylogenetic inference, Sober provides some biological arguments for the following theoretical assumptions:

(i) *Intermediate Probabilities*. All probabilities are strictly greater than 0 and less than 1.

(ii) *Conditional Independence*. A common cause screens off one joint effect from another, and a more proximal cause of an effect screens off a less proximal cause from the effect. That is, a tree is Markovian and singly connected, (Sober 1984, p.224). In our case this condition is that for all \(i, j,\) and \(m:\)

\[
P(A = i.B = j/X = m) = P(A = i/X = m).P(B = j/X = m),
\]

where \(X\) stands for either the variable \(D\) or \(E\) in that case of \((AB)C\) and \(X = E\) for \((ABC).\) Similar conditions hold for \(P(A = i.C = k/E = m)\) and \(P(B = j.C = k/E = m).\)

(iii) *The Required Inequality*. For all transition probabilities of the form \(P(X = i/Y = m)\) from past events to later events, we have that

\[
P(X = 1/Y = 1) - P(X = 1/Y = 0) > 0.
\]
For example, this requires that for hypothesis \((AB)C\), \(P(A = 1/D = 1) - P(A = 1/D = 0) > 0\), and \(P(D = 1/E = 1) - P(D = 1/E = 0) > 0\) (from which it follows that \(P(A = 1/E = 1) - P(A = 1/E = 0) > 0\), so the requirement is consistent). In the language of section 2, there is a positive statistical correlation between any two variables linearly connected in the tree.

(iv) Primitive Ancestry. The character state at the root of all trees is always in the primitive state. That is, \(E = 0\).

The importance of these theoretical postulates is due to the constraints they place on the predicted probabilities of observed occurrences. Let \(P_0\) denote an arbitrary probability function compatible with the tree \((ABC)\) as dictated by the theoretical constraints (i), (ii), (iii) and (iv). There will be many such probability functions because assumptions (i) to (iv) are not strong enough to specify the probability uniquely (the hypothesis is incomplete in this sense). Then we can prove that (Forster 1986b) for any such \(P_0\), for all \(i, j, and k \in \{0, 1\},\)

\[
P_0(A = i, B = j, C = k) = P_0(A = i).P_0(B = j).P_0(C = k).
\]

Similarly, if we let \(P_1\) and \(P_2\) denote arbitrary probability functions compatible with \((AB)C\) and \(A(BC)\) respectively, via constraints (i), (ii), (iii), and (iv), then we can prove that

\[
P_1(A = i, B = j, C = k) = P_1(A = i, B = j).P_1(C = k) \tag{7}
\]

\[
P_1(A = 1, B = 1) - P_1(A = 1).P_1(B = 1) > 0 \tag{8}
\]

\[
P_2(A = i, B = j, C = k) = P_2(A = i).P_2(B = j).P_2(C = k) \tag{9}
\]

\[
P_2(B = 1, C = 1) - P_2(B = 1).P_2(C = 1) > 0. \tag{10}
\]

The constraints (6), (7) and (9) arise from the assumption that \(E = 0\), while (8) and (10) follow mainly from the assumption of positive correlation, (iii). But still these constraints are not sufficient to provide a likelihood comparison of \((ABC)\), \((AB)C\), and \(A(BC)\) because the probability values \(P_0(A = i, B = j, C = k)\), \(P_1(A = i, B = j, C = k)\), and \(P_2(A = i, B = j, C = k)\) have not been given unique numerical values.

Beyond the theoretical constraints on those functions, there is no evidence for the actual values for the transition probabilities, and thus no specification of the probabilities of \((A = i, B = j, C = k)\). So, no likelihood comparison can be made of these purely structural hypotheses. Each hypothesis, in other words, is actually represented by a set of probability functions satisfying their respective constraints—a set \(\{P_0\}\) for \((ABC)\), a set \(\{P_1\}\) for \((AB)C\) and a set \(\{P_2\}\) for \(A(BC)\). Moreover, it is typical of the problem of phylogenetic inference that there are insufficient theoreti-
Figure 2. Sober’s cross-identification of paths for the hypotheses \((ABC), (AB)C\), and \(A(BC)\) respectively.

ical or empirical restrictions available to specify uniquely a probability function.

Sober puts forward an interesting suggestion about how to bypass this problem. His idea is to compare sets of probability functions by making a “natural” mapping between their members and then comparing each member with its image under the mapping in terms of likelihood. If such a comparison is universally in favor of the members of one set over the other, Sober then says we have a likelihood justification of the structural hypothesis represented by that set.

In the example of phylogenetic inference, this program is implemented by first cross-identifying paths on different trees as drawn in Figure 2. The transition probabilities, though unknown, are the same for each path no matter which hypothesis we refer to. Thus, we have:

\[ P_0(X = 1/Y = 1) = P_1(X = 1/Y = 1) = P_2(X = 1/Y = 1) = q_i, \]
\[ P_0(X = 1/Y = 0) = P_1(X = 1/Y = 0) = P_2(X = 1/Y = 0) = e_i. \]

For example, for path 2;

\[ P_0(B = 1/E = 1) = P_1(B = 1/D = 1) = P_2(B = 1/D = 1) = q_2, \]
\[ P_0(B = 1/E = 0) = P_1(B = 1/D = 0) = P_2(B = 1/D = 0) = e_2. \]

(Note that Sober’s assumption (iii) can now be written more concisely as \(q_i > e_i\) for all paths \(i\).)

This cross-identification of transition probabilities induces a mapping (though not one-to-one) between the sets \(\{P_0\}, \{P_1\}, \) and \(\{P_2\}, \) since each probability function in these sets is determined by given values of the transition probabilities (given the assumptions (i)–(iv)). That is, a given set of values for the \(q_i\)'s and \(e_i\)'s will determine (unique) probability values \(P_0(A = i,B = j,C = k), P_1(A = i,B = j,C = k), \) and \(P_2(A = i,\)
$B = j. C = k$), and these correspond to each other under the induced mapping. Writing $P_1(110)$ as shorthand for $P_1(A = 1. B = 1. C = 0)$, etc., the important results, as stated by Sober (1984, pp. 224–225), are: For all possible values of $q_1, e_1, \ldots, q_4, e_4$;

\begin{align}
P_1(110) &> P_2(110), \\ P_1(110) &> P_0(110), \\ P_1(001) &< P_2(001), \\ P_1(001) &< P_0(001).
\end{align}

In words, the hypothesis $(AB)C$ has greater likelihood than either $(ABC)$ or $A(BC)$ relative to the data $(A = 1, B = 1, C = 0)$, whereas $(AB)C$ has the worst likelihood relative to the data $(A = 0, B = 0, C = 1)$. So the best explanation of the matching $(A = 1, B = 1)$ is by positing a recent common cause between those two events, but not so for the matching $(A = 0, B = 0)$!

The latter result is Sober’s alleged counterexample to the principle of common cause, which he reads as saying that matchings should not always be explained by common causes (in this case the 0–0 matching should not). First note, contra Sober, that this is not a counterexample to any common cause principle that takes its explanandum to be a statistical correlation, but I believe that such a counterexample could be constructed if we were to accept Sober’s likelihood comparisons in (13) and (14). In any case, I will dispute those two results.

Sober’s conclusion depends crucially on his CIP assumption (v). But the mapping of transition probabilities defined there has some undesirable properties, and this undermines his argument. In particular, the mapping is not invariant under extensions of the models. Both $(AB)C$ and $A(BC)$ provide extendible models of the phylogenetic history of the three populations $A, B$ and $C$ in the sense that their taxonomic relationship with other populations might also be considered. Suppose we include one other population, $F$, in the picture. There are at least three different ways of extending the hypothesis $(AB)C$ to include $F$. These are $((AF)B)C$, $A(BF)C$, and $A(B)CF$. Similarly, $A(BC)$ may be extended to $(AF)(BC)$, $A((BF)C)$ or $A(B(FC))$. Conversely, if we choose to ignore $F$, the first three hypotheses reduce to $(AB)C$ and the last three to $A(BC)$.

Now suppose we compare hypothesis $(AB)(CF)$ with $(AF)(BC)$ as drawn in Figures 3 and 4 respectively.

From Sober’s assumptions, (i), (ii), (iii) and (iv) stated earlier, the con-
strains on any probability function \( P_3 \) conforming to \((AB)(FC)\), and \( P_4 \) conforming to \((AF)(BC)\) are, respectively:

\[
\]

\[
\]

If we now choose to ignore \( F \), then \((AB)(FC)\) reduces to \((AB)C\) and \((AF)(BC)\) reduces to \( A(BC) \), and indeed the constraint (15) reduces to (7) and constraint (16) reduces to (9) if we eliminate \( F \) by summing over \( l \), as we would hope. For example,

\[
\]

\[
+ P_3(A = i.B = j.C = k.F = 0)
\]

\[
= P_3(A = i.B = j)P_3(C = k.F = 1)
\]

\[
+ P_3(A = i.B = j)P_3(C = k.F = 0)
\]

\[
= P_3(A = i.B = j)[P_3(C = k.F = 1)
\]

\[
+ P_3(C = k.F = 0)]
\]

\[
= P_3(A = i.B = j)P_3(C = k).
\]

Thus, there is a natural reduction of the constraints for \(((AF)B)C\), \((A(BF)C)\) and \((AB)(FC)\) to the constraint for \((AB)C\) and those for \((AF)(BC)\), \((A(BF)C)\) and \((A(BC))\) to that for \( A(BC) \) when we eliminate \( F \). So, the constraints derivable from assumptions (i)--(iv) have the desirable property that they are invariant under extensions of the model in the sense described. But Sober’s assumption (v) does not fulfill this requirement because the path numbering in Figures 3 and 4 does not reduce the path labeling in Figure 2 when we ignore \( F \). This has the detrimental
consequence that the ordering between \( P_3(001) \) and \( P_4(001) \) is exactly the reverse of that between \( P_1(001) \) and \( P_2(001) \). In particular,
\[
P_3(001) = [(1 - e_5)(1 - e_4)(1 - e_2) \\
+ e_6(1 - q_1)(1 - q_2)][(1 - e_5)e_4 + e_5q_4] \\
P_4(001) = [(1 - e_6)(1 - e_1) \\
+ e_6(1 - q_1)][(1 - e_5)(1 - e_2)e_4 + e_5(1 - q_2)q_4]
\] (17) (18)
from which we can prove that
\[
P_3(001) - P_4(001) = [(1 - q_2) \\
- (1 - e_2)][e_6(1 - e_5)(1 - q_1)e_4 - (1 - e_6)e_5(1 - e_1)q_4].
\] (19)
If we make the simplifying assumption that \( e_6 = e_5 \), we can prove that \( P_3(001) \) is strictly greater than \( P_4(001) \), whereas \( P_1(001) \) is invariably less than the corresponding \( P_2(001) \). This shows that the correspondence set up between \( P_3 \)'s and \( P_4 \)'s has different properties entirely. Hence, the mapping set up between the probability functions in \((AB)C\) and \(A(BC)\) is not invariant under extensions of the models, in the sense defined, and any conclusion based upon such a mapping is suspect. In particular, we can easily arrive at the conclusion that a 0–0 matching between \( A \) and \( B \) is positive evidence of a recent common cause between those events, looking at the situation in Figures 3 and 4, contrary to the conclusion arrived at by the cross-identification of paths in Figure 2. The conclusion to be drawn from this is that Sober's strategy of identifying paths between trees does not solve the problem of how to compare incomplete structural hypotheses. So, what alternative will work? How should we measure the evidential support of incomplete hypotheses?

5. The Best-of-the-Best Strategy. The structural hypotheses \((ABC)\), \(((AB)C)\), and \((A(BC))\), are represented by sets of probability functions \( \{P_0\} \), \( \{P_1\} \), and \( \{P_2\} \), respectively. A straightforward way of bypassing the problems inherent in Sober's assumption (v) is the following. Choose the probability function from each set with the highest likelihood and then compare the likelihood of these. That is, we take the best of each set and then choose the best of these, so I will call this the "best-of-the-best" method. This strategy clearly requires no mapping between probability sets, and so makes no use of assumption (v) or anything like it. (The method is unlikely to be new, but I personally do not know of any explicit discussions of it.)

In application to Sober's examples, the best-of-the-best method of likelihood comparison does not make any decisive judgment on the evidence
of single cases, because a single instance will give, for example, \( r(110) = 1 \), and \( P_0(110) \), \( P_1(110) \), and \( P_2(110) \) can all be made to be arbitrarily close to 1 under their respective constraints. And the same applies to \( P_0(001) \), \( P_1(001) \), and \( P_2(001) \) if we observe that \( r(001) = 1 \). So, according to this strategy, no decision is reached by comparing likelihoods relative to single observations of matchings between events. On the other hand, when our data consist of several observations then the best-of-the-best method does give unambiguous results.

Thus, we have returned to the point of disagreement over whether common cause hypotheses should explain the matching of event tokens or the observed statistical correlations between variables. Let us consider one of Sober's own examples. He supposes that we observe 10 characteristics in three species \( A \), \( B \), and \( C \) to obtain the following table of results.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( B )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( C )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

With the added assumptions that the same probability distribution applies to each characteristic, which are taken to be mutually independent, it is easily shown that \((ABC)\) is marginally the best hypothesis (see Forster 1986b, for some more decisive examples). The reason is simple enough to understand. \((AB)C\) and \(A(BC)\) predict a positive correlation \(\rho(A,B)\) and \(\rho(B,C)\) respectively (see inequalities (8) and (10)), whereas none is observed (the observed covariances are 0). To use an analogy, \((ABC)\) best models the situation in terms of independent random draws from three urns \( A \), \( B \), and \( C \), where 90% of the marbles in urn \( A \), 100% of marbles in urn \( B \), and 10% of marbles from urn \( C \) are labeled with a "1". The best-of-the-best method for comparing incomplete hypotheses confirms the intuition that there is no reason to reject the independence hypothesis \((ABC)\) as providing the best explanation. The high number of 1–1 matchings between \( A \) and \( B \) is adequately explained by supposing there are high proportions of those marbles in both urns.

This conclusion is contrary to Sober's conclusion, which favors \((AB)C\) in this case on the basis that the majority of 1–1 matchings occur between \( A \) and \( B \). On Sober's view, the hypothesis \((ABC)\) fares even worse than \(A(BC)\) because it does not posit any common causes to explain any of the 1–1 matchings. There are two mistakes made by Sober here. The first is the philosophical mistake of thinking that each single 1–1 matching is best explained in terms of a common cause. This is the mistake, pointed out by Aristotle, of trying to explain why two men meet at the market.
There is no common cause explanation required in this case. The second mistake is mathematical. Sober tacitly assumes that, when several characteristics are observed, the total weight of evidence in favor of any hypothesis combines the weight of evidence from each single characteristic additively; in other words, he seems to assume that the likelihood relative to the total evidence is equal to the sum of the likelihoods relative to each single observation. In fact, likelihoods combine multiplicatively; that is, the likelihood relative to total evidence is the product of the likelihood relative to each single observation (assuming that the observations are mutually independent).

It is a general fault of Sober’s method (with likelihoods correctly combined multiplicatively) that it will usually remain undecided in the face of large numbers of observations because the universality condition will seldom hold—it will seldom be the case that all probability functions in one set will have higher likelihoods than their counterparts in a second set. Moreover, whenever the best-of-the-best method makes a positive judgment in favor of one hypothesis over another, Sober’s method must either remain neutral or agree with its judgment (the former being more likely). The argument for this is as follows: suppose the maximum likelihood within the set \( \{P_1\} \) is greater than the maximum in \( \{P_2\} \). Sober’s method defines a mapping between the sets \( \{P_1\} \) and \( \{P_2\} \) such that every element of \( \{P_1\} \) is mapped to some element of \( \{P_2\} \). Consider the “best” element in \( \{P_1\} \). It is better than the element in \( \{P_2\} \) to which it is mapped because it is better than all members of \( \{P_2\} \), by hypothesis. So, it cannot be the case that all elements of \( \{P_2\} \) have greater likelihoods than their images in \( \{P_1\} \) under Sober’s mapping, or any mapping for that matter. It is only in the odd case in which the best-of-the-best strategy makes no judgment (that is, remains neutral) that Sober’s method can contradict it, as is the case for single observations.

We saw at the end of section 4 that Sober’s method of likelihood comparisons does not work. But how do we go about showing that the best-of-the-best strategy does better? The proof of the pudding is usually in the eating. Reasons in favor of the best-of-the-best strategy should be obtained by examining its consequences in a wide variety of applications and seeing if they accord with accepted practice. In the next section I will examine the consequences of this method for phylogenetic inference. Other applications will appear in future papers.

But there are also some direct arguments in its favor. First, it is worth noting that the best-of-the-best strategy is a rather natural extension of the maximum likelihood method of estimating the parameter values of curves fitted to random data. Suppose that we have data on the inter-relationship of two continuous random variables \( A \) and \( B \) displayed on a scatter diagram. Now consider two (incomplete) hypotheses giving the
functional form of the relationship as:

\[ A = f(B; a, b) = a + b \cdot B \]  

(20)

\[ A = g(B; c, d, e) = c + d \cdot B + e \cdot B^2 \]  

(21)

Both of these hypotheses are incomplete in the sense that the parameters \( a, b, c, d, \) and \( e \) are unspecified. Thus, they can be represented by the sets of functions \( \{f(B; a, b): a \in \mathbb{R}, b \in \mathbb{R}\} \) and \( \{g(B; c, d, e): c \in \mathbb{R}, d \in \mathbb{R}, e \in \mathbb{R}\} \) respectively. Sober’s strategy would be to set up some mapping between the two sets (for example, by putting \( a = c \) and \( b = d \)) and then comparing each function with its image under the mapping. The immediate disadvantage of this strategy is that it will very rarely, if ever, produce an unambiguous result. Suppose that the data is in fact very closely clustered around a parabolic curve. Then we would want (21) to win out over (20) (that is, in terms of evidential support—considerations of simplicity, etc., may alter our final choice). But I fail to see how Sober’s method could produce this result, because for some values of parameters \( a \) and \( b \) adding a term \( e \cdot B^2 \) will worsen the fit, not improve it. (The same problem occurs for Sober’s own example of phylogenetic inference as argued earlier.)

The obvious solution would be to find maximum likelihood estimations of all the parameters, and then compare the fit of those two curves.\(^4\) That is, we can first find the best fitting function of the form \( f(B; a, b) \) and the best fitting function of the form \( g(B; c, d, e) \), and then choose between them on the basis of likelihood. What I have called the best-of-the-best method is completely analogous, except that the parameters in the evolutionary models (the transition probabilities) are not uniquely determined by the maximum likelihood procedure. But the basic aim is to accomplish an effective theory comparison, and this is achieved.

There is another methodological disadvantage of Sober’s strategy. As Sober will agree, evidential support is not the only factor operative in accepting hypotheses. Often, there is a tradeoff between the level of evidential support and other factors such as simplicity. But to operate any such tradeoff we need some sort of \textit{quantitative} measure of the difference in support between two hypotheses. For instance, Edwards (1972) takes a difference in support (the logarithm of the likelihood) of about 2 as a reasonable cutoff for rejecting the “null” hypothesis, showing that this corresponds to a 5% probability of rejecting a true null hypothesis (Type I error) under normal conditions. But Sober’s method of likelihood com-

\(^4\)If the error distribution of \( A \) given \( B \) is normal with the same variance about its mean for every value of \( B \), then the maximum likelihood estimation of the parameters reduces to the well-known method of least squares.
comparison—unless mine—allows for no such quantitative comparison of incomplete theories.

6. Consequences for Phylogenetic Inference. If we adopt Sober’s theoretical framework (assumptions (i) to (iv)), then the genealogical hypotheses \((ABC)\), \((AB)C\), and \(A(BC)\) are represented by sets of probability functions \(\{P_0\}\), \(\{P_1\}\), and \(\{P_2\}\), respectively, which satisfy the constraints (6) to (10). Our data consist of the observations of the characteristic states of species \(A\), \(B\), and \(C\) for many different characteristics. Suppose that the observed relative frequencies for some particular values of \(i\), \(j\) and \(k\) (equal to 0 or 1) are given as \(r(A = i, B = j, C = k)\); written \(r_{ijk}\).\(^5\) (Let us also denote \(r(A = i, B = j)\) by \(r_{ij}\), \(r(B = j, C = k)\) by \(r_{jk}\), and \(r(A = i)\) by \(r_i\), etc.). The relevant feature of the observed relative frequencies will be the values of the covariances \(r(A = 1, B = 1) - r(A = 1)r(B = 1)\) and \(r(B = 1, C = 1) - r(B = 1)r(C = 1)\) denoted by \(C_{AB}\) and \(C_{BC}\) respectively.

Extending the terminology of Edwards (1972), we define the support of any incomplete hypothesis as the logarithm of the maximum likelihood value under the constraints imposed (which is the same as the maximum support value), in accordance with the idea behind the best-of-the-best method of likelihood comparisons. For example, the value of Support\((ABC)\) is the maximum value of \(\log P_1(\text{data})\), where \(P_1(\text{data})\) is shorthand notation for the probability of the observations given according to the probability function \(P_1\), and where \(P_1\) ranges over the members of the set \(\{P_1\}\). There is a mathematical theorem (see Forster 1986b, for details) that tells us that this maximum occurs when the theoretical probabilities given by \(P_1\) match the observed relative frequencies as “closely” as possible. The maximum obtained for our phylogenetic hypotheses are given as functions of the observed relative frequencies, as follows:

\[
\text{Support}(ABC) = N \sum_i \sum_j \sum_k r_{ijk} \log r_{ij}r_{jk} \quad (22)
\]

\[
\text{Support}((AB)C) = N \sum_i \sum_j \sum_k r_{ijk} \log r_{ij}r_{jk}, \quad \text{if} \quad C_{AB} \leq 0 \quad (23)
\]

\[
\text{Support}((AB)C) = N \sum_i \sum_j \sum_k r_{ijk} \log r_{ij}r_{jk}, \quad \text{if} \quad C_{AB} > 0 \quad (24)
\]

\(^5\)More pedantically, we should index each variable to the characteristic in question and take \(A_n = i\) to mean that species \(A\) is in state \(i\) for the \(n\)th characteristic. The relative frequency \(r(A = i)\) is then shorthand notation for \(1/N \sum_n N(A_n = i)\), where \(N\) is the total number of characteristics observed, and \(N(A_n = i)\) is the characteristic function whose value is 1 if \(A_n = i\) is true, and 0 otherwise.
\[ \text{Support}(A(BC)) = N \sum_{i} \sum_{j} \sum_{k} r_{ijk} \log r_{ij} r_{ik}, \quad \text{if} \quad C_{BC} \leq 0 \quad (25) \]

\[ \text{Support}(A(BC)) = N \sum_{i} \sum_{j} \sum_{k} r_{ijk} \log r_{jk}, \quad \text{if} \quad C_{BC} > 0 \quad (26) \]

We will now assume that we have non-trivial relative frequencies in the sense that \( r(A = 1) \), \( r(B = 1) \), and \( r(C = 1) \) are not 0 or 1. (This eliminates a single observation from consideration—we will return to that case later.) There are four possible types of data sets, which we consider in turn.

**Case 1.** \( C_{AB} \leq 0 \), and \( C_{BC} \leq 0 \): Each of \((ABC)\), \((AB)C\), and \(A(BC)\), will compare equally (badly) with the data and the support for each will the same.

**Case 2.** \( C_{AB} > 0 \), and \( C_{BC} \leq 0 \): \(A(BC)\) will have the same support as \((ABC)\), but \((AB)C\) will receive greater support than both.

**Case 3.** \( C_{AB} \leq 0 \), and \( C_{BC} > 0 \): \((AB)C\) will have the same support as \((ABC)\), but \(A(BC)\) will receive greater support than both.

**Case 4.** \( C_{AB} > 0 \), and \( C_{BC} > 0 \): \((ABC)\) will receive less support than either \((AB)C\) or \(A(BC)\), but the comparison between \((AB)C\) and \(A(BC)\) is more difficult. In the special case in which \( r(A = 1) = r(B = 1) \); \( \text{Support}((AB)C) > \text{Support}(A(BC)) \) if and only if, \( C_{AB} > C_{BC} \). In the general case when \( r(A = 1) \neq r(C = 1) \), \((AB)C\) will still be favored over \(A(BC)\) when \( C_{AB} \) is sufficiently greater than \( C_{BC} \), and vice versa. (See Forster 1986b, for detailed proofs.)

Intuitively, these results say nothing more than that hypothesis \((AB)C\), for instance, best fits the data when there is positive correlation observed between \(A\) and \(B\) because that is what the hypothesis predicts should be the case (most of the time). But it is only the prediction about statistical correlations that matters; other features of the relative frequencies do not play a role in comparing hypotheses. The observed correlations (as measured by the covariances \(C_{AB}\) and \(C_{BC}\)) function as the relevant statistics for deciding which causal hypothesis is best supported by the evidence. This is a consequence of the best-of-the-best method for comparing incomplete theories. Causal hypotheses are only compared by their predictions concerning statistical correlations and nothing more. In consequence, causal hypotheses only explain statistical correlations, just as in Reichenbach’s original schema. When \(A\) and \(B\) randomly covary, as in the example of drawing marbles independently from different urns, the
value of $C_{AB}$ is close to 0; so there is nothing that needs explaining in terms of common causes. Indeed, the common cause hypothesis (CC) has significantly greater support over its single cause rival (SC)\(^6\) only if the observed statistical correlation between $A$ and $B$ is sufficiently greater than 0. Therefore, the present likelihood method provides a straightforward likelihood rationale for Reichenbach’s original schema and shows how it extends to the example of phylogenetic inference as well.

7. Concluding Remarks. The main point of this paper is to show that Sober’s (1984) ideas on common cause explanation do not get off the ground. Causal explanations are generally incomplete in that they do not specify their own initial conditions (nor the probabilities thereof), so they do not uniquely specify the theoretical probability of the data. Therefore, no likelihood measure of evidential support can be given without further methodological rules. Sober’s own method fails because the mapping between probability sets is altogether artificial, whereas my own proposal of defining the support of the set as the maximum support of its elements avoids this sort of difficulty, as well as being closer to standard statistical methods.

Sober views common causes as explaining matchings between events, but the argument for this was in terms of Sober’s method of comparing incomplete theories, and this has been severely damaged. Of course, it does not follow that his conclusion is false just because an argument for it is invalid. But, it is false if we accept the best-of-the-best method of defining the support of incomplete theories. For then no single observation by itself can support one causal hypothesis over another, and conversely no causal hypothesis is designed to explain the matchings of token events.

However, it is possible to ask a slightly different question. What is the contribution of a single observation within the context of a larger data set. It can be shown (Forster 1986b) that, in Sober’s framework, a 110 observation cannot give worse support to $(AB)C$ than it does to $(ABC)$ or $A(BC)$ in the precise sense that

$$P_1(110) \geq P_0(110) \quad \text{and} \quad P_1(110) \geq P_2(110). \quad (27)$$

This partly recovers Sober’s results $P_1(110) \geq P_0(110)$ and $P_1(110) \geq P_2(110)$ (inequalities (11) and (12)). But his other results $P_1(001) < P_2(001)$

\(^6\)We assume here that the separate cause hypothesis posits independent causes for $A$ and $B$, for otherwise the SC hypothesis will be empirically indistinguishable from the CC hypothesis (within the present framework).
and \( P_1(001) < P_0(001) \) (inequalities (13) and (14)) are completely overturned by the result that

\[
P_1(001) \succeq P_0(001) \quad \text{and} \quad P_1(001) \succeq P_2(001).
\]

(28)

So, any kind of matching, whether 1–1 or 0–0, is positive evidence of a common cause in this sense. (Note that this depends on the theoretical correlations being positive, as stated in assumption (iii).) The symmetry between 0–0 and 1–1 matchings is due to the fact that the covariances, \( C_{AB} \) etc., are the relevant statistics and they have the property that

\[
C_{AB} = r(A = 1, B = 1) - r(A = 1)r(B = 1)
= r(A = 0, B = 0) - r(A = 0)r(B = 0).
\]

(29)

The more 1–1 matchings above the proportion expected by chance \((r(A = 1)r(B = 1))\) the higher the value of \( C_{AB} \) and the better supported is the CC explanation. But exactly the same applies with respect to 0–0 matchings. If Sober had provided a convincing argument for his inequalities (13) and (14), he may have had not only a counterexample to Reichenbach’s schema, but also an argument against viewing common causes as explaining correlations in the statistical sense. But Sober’s argument is flawed.

There is another example that Sober uses to support the same conclusion, so I must refute this as well. Sober argues as follows:

The 001 observation provided an example of a matching that is not best treated via a CC explanation. The lamp case may be modified to yield a case of this sort. Suppose that your two lights A and B are on, but that your neighbor’s lamp C is not. It is known that all three are connected to the power station that is sending electricity, and that its sending electricity is a positive causal factor in producing lit lamps. In this circumstance, the matching of your two lamps is not best explained by postulating a common cause unique to them. There may, of course, be other reasons for thinking that the two lamps are on in virtue of there being a common cause that affects them but not the lamp across the street. But the minimum assumptions made in our likelihood argument do not imply this. (Sober 1984, p. 234)

I agree that there is little evidence that my two lamps are controlled by a common switch on this evidence. However, Sober’s own likelihood method forces him to a stronger conclusion, namely that this evidence better supports \( A(BC) \) than \( (AB)C \). But the idea that it is more likely that

\text{This result is stated by Sober himself (1984, p. 225) and is equivalent to equation (13) of this paper. Of course, the objection assumes that the lamp example satisfies Sober’s theoretical constraints, (i) to (iv) in section 4. But I gather from the passage quoted above that Sober believes this to be the case.
one of my lamps shares a common control switch with my neighbor’s lamp is most disturbing. There is definitely no intuitive support for this conclusion.

Admittedly, there is a justified intuition that my lamps share a common switch if we observe them simultaneously going off, or coming on. But the example now supports the Reichenbachian view of the matter, because in observing a change we have actually observed several matchings rather than a single joint event. At times before $t_0$ we observe a series of 0–0 matchings, and at times after $t_0$ we observe a series of 1–1 matchings. When we collect these observations together we have a strong statistical correlation, and this is what proves the common cause hypothesis. And it is the statistical correlation, and not the event correlations that the hypothesis explains.

REFERENCES