Are There Algorithms That Discover Causal Structure? 30 June 1998

David Freedman, Department of Statistics,
University of California, Berkeley, CA 94705

Paul Humphreys, Department of Philosophy,
University of Virginia, Charlottesville, VA 22903

Technical Report 514, Department of Statistics
University of California, Berkeley, CA 94720

Abstract

For nearly a century, investigators in the social and life sciences have used regression models to deduce cause-and-effect relationships from patterns of association. Path models and automated search procedures are more recent developments. However, these formal procedures tend to neglect the difficulties in establishing causal relations, and the mathematical complexities tend to obscure rather than clarify the assumptions on which the analysis is based. This paper focuses on statistical procedures that seem to convert association into causation. Formal statistical inference is, by its nature, conditional. If maintained hypotheses A, B, C, ... hold, then H can be tested against the data. However, if A, B, C, ... remain in doubt, so must inferences about H. Careful scrutiny of maintained hypotheses should therefore be a critical part of empirical work—a principle honored more often in the breach than the observance.

Spirtes, Glymour, and Scheines have developed algorithms for causal discovery. We have been quite critical of their work. Korb and Wallace, as well as SGS, have tried to answer the criticisms. This paper will continue the discussion. Their responses may lead to progress in clarifying assumptions behind the methods, but there is little progress in demonstrating that the assumptions hold true for any real applications. The mathematical theory may be of some interest, but claims to have developed a rigorous engine for inferring causation from association are premature at best. The theorems have no implications for samples of any realistic size. Furthermore, examples used to illustrate the algorithms are diagnostic of failure rather than success. There remains a wide gap between association and causation.

1. Introduction

In Humphreys and Freedman (1996), we showed that the program of automated causal inference described in *Causation, Prediction, and Search* by Spirtes, Glymour and Scheines (SGS, 1993) is seriously—even fatally—flawed. Here, we put our arguments into a broader context and reply to the comments of Korb and Wallace (1997) and SGS (1997). To make the present paper relatively self-contained, we describe the SGS program and the critique in Section 1, along the lines of our earlier publications.[1] Section 2 replies to Korb and Wallace, while Section 3 replies to SGS.

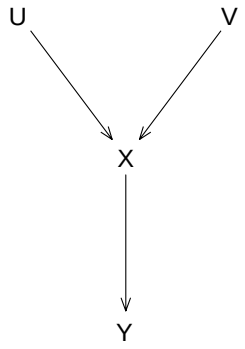
We believe, along with many others, that identifying causal relations requires thoughtful, complex, unrelenting hard work; substantive scientific knowledge plays a crucial role. Claims to have automated that process require searching examination. The principal ideas behind automated causal inference programs are hidden by layers of formal technique. Therefore, it is important to make the ideas explicit and probe them carefully. SGS illustrate the problem; these authors contend they have algorithms for discovering causal relations based only on empirical data, with no little or no need for subject-matter knowledge. Their methods—which combine graph theory, statistics and computer science—are supposed to allow quick, virtually automated conversion of statistical association to causation. Their algorithms are held out as superior to methods already in use in the social sciences (regression analysis, path models, factor analysis, hierarchical linear models, and so on). According to SGS, researchers who use these other methods are sometimes too timid, sometimes too bold, and often just misguided.

> Chapters 5 and 8 illustrate a variety of cases in which features of linear models that have been justified at length on theoretical grounds are produced immediately from empirical covariances by the procedures we describe. We also describe cases in which the algorithms produce plausible alternative models that show various conclusions in the social scientific literature to be unsupported by the data. [SGS, 1993, p. 14]
>
> In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled studies. So far as we can tell, the popular automatic regression search procedures [like stepwise regression] should not be used at all in contexts where causal inferences are at stake. Such contexts require improved versions of algorithms like those described here to select those variables whose influence on an outcome can be reliably estimated by regression. [SGS, 1993, p. 257]

Such claims are quite exaggerated, and in fact there are no real examples where the algorithms succeed. The algorithms themselves may well be of some interest, but the technical apparatus is only tangentially related to long-standing philosophical questions about the meaning of causation, or to real problems of statistical inference from imperfect data. This section will summarize the evidence.

First, we sketch the idea of path models and the SGS discovery algorithms. Statistical relationships are often displayed in graphical form, path models being one example. These models represent variables as nodes in a graph. An arrow from X to Y means that X is related to Y, given the prior variables.[2] In Figure 1, for example, the regression equation for Y in terms of U, V, and X should include only X: the only arrow into Y is from X. However, the equation for X in terms of U and V should include both variables: there are arrows into X from U and V.

Figure 1. Directed Acyclic Graphs.



Two of the central ideas behind the SGS discovery algorithms are the 'Markov condition' for graphs, developed by Kiiveri and Speed (1982, 1986), and the 'faithfulness assumption' due to Pearl (1988).[3] These are purely mathematical assumptions relating graphs and probabilistic independence. (There will be an informal discussion of these assumptions, below.) SGS focus on a special class of graphical models, the Directed Acyclic Graphs (DAGs). Properties of these graphs are summarized in SGS (1993, Chapter 2); the Markov condition and the faithfulness assumption are also stated there. Starting from the joint distribution of the variables, the Markov condition, and the faithfulness assumption, SGS have algorithms for determining the presence or absence of arrows.

The Markov condition says, roughly, that certain nodes in the graph are conditionally independent of other nodes, where independence is a probabilistic concept. (More precisely, a node in the graph stands for a random variable, and it is the variables that may or may not be independent.) In Figure 1, for example, Y is independent of U and V given X. With DAGs, there is a mathematical theory that permits conditional independence relations to be read off the graph. The faithfulness assumption says there are no 'accidental' relations: conditional probabilistic independence holds according to presence or absence of arrows, not in virtue of specific parameter values. Under such circumstances, the probability distribution is said to be 'faithful' to the graph.[4] If the probability distribution is faithful to a graph for which the Markov condition holds, that graph can be inferred (in whole or in part) from conditional independence relations, and the object of the SGS algorithms is to reconstruct the graph from independence relations.

There is no coherent ground—just based on the mathematics—for thinking that the graphs represent causation. The connection between arrows and causes is made on the basis of yet another assumption, the 'causal Markov condition' (SGS, 1993, Chapter 3). Moreover, according to the 'causal representation convention' (p. 47), causal graphs are DAGs where arrows represent causation. In short, the causal Markov condition is just the Markov condition, plus the assumption that arrows represent causation. Thus, causation is not a consequence of the theory, it is just another assumption. To compound the confusion, SGS also make the convention (p. 56) that the 'Markov property' means the 'causal Markov property'.

Formal theories nowadays either use uninterpreted formulas as axioms, or define classes of abstract structures by the axiomatization itself. These axiomatic approaches

make a clear distinction between a mathematical theory and its interpretations. SGS do not use these approaches, and positively invite the confusion that axiomatics are supposed to prevent. Indeed, SGS seem to have no real concerns about interpretative issues:

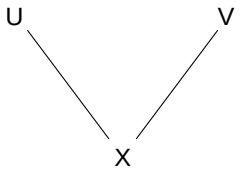
> Views about the nature of causation divide very roughly into those that analyze causal influence as some sort of probabilistic relation, those that analyze causal influence as some sort of counterfactual relation (sometimes a counterfactual relation having to do with manipulations or interventions), and those that prefer not to talk of causation at all. We advocate no definition of causation, but in this chapter attempt to make our usage systematic, and to make explicit our assumptions connecting causal structure with probability, counterfactuals and manipulations. With suitable metaphysical gyrations the assumptions could be endorsed from any of these points of view, perhaps including even the last. [p. 41][5]

SGS do not give a reductive definition of 'A causes B' in non-causal terms. And their apparatus requires that you already understand what causes are. Indeed, the causal Markov condition and the faithfulness assumption boil down to this: causes can be represented by arrows when the data are faithful to the true causal graph that generates the data. Thus, causation is defined in terms of causation, with little value added.[6] The mathematics in SGS will not be of much interest to philosophers seeking to clarify the meaning of causality.

The SGS algorithms for inferring causal relations from data are embodied in a computer program called TETRAD II. We give a rough description. The program takes as input the joint distribution of the variables, and it searches over DAGs. In real applications, of course, the full joint distribution is unknown, and must be estimated from sample data. In its present incarnation, TETRAD can handle only two kinds of sample data, governed by conventional and unrealistic textbook models: (i) independent, identically distributed multivariate Gaussian observations, or, (ii) independent, identically distributed multinomial observations. These assumptions are not emphasized in SGS, but appear in the computer documentation and the computer output.[7]

In essence, TETRAD begins with a 'saturated' graph, where any pair of nodes are joined by an edge. If the null hypothesis of independence cannot be rejected—at, say, the 5% level, using some variation on the t-test—the edge is deleted. The statistical test is relevant only because of the statistical assumptions. After examining all pairs of nodes, TETRAD moves on to triples, and so forth. According to the faithfulness assumption, independence cannot be due to the cancellation of conditional dependencies. That is why an edge, once deleted, never returns.

Figure 2. Orienting the edges.



TETRAD also orients the edges left in the graph. (Orienting an edge between two variables says which is the cause and which the effect.) Take the graph in Figure 2. If

U and V are conditionally independent given X, the arrows cannot go from U and V to X—that would violate the faithfulness assumption: thus, U and V are the effects, X is the cause. However, it is exact independence that is relevant, and exact independence cannot be determined from any finite amount of sample data. Consequently, the mathematical demonstrations in SGS (e.g., Theorem 5.1 on p. 405) do not cope with the most basic of statistical ideas. Even if all the assumptions hold, statistical tests make mistakes. The tests have to make mistakes, because sample data do not determine the joint distribution. The problem is compounded when, as here, multiple tests are made. Therefore, the SGS algorithms can be shown to work only when the exact conditional independencies and dependencies are given. Similarly, with the faithfulness condition, it is only exact conditional independence that protects against confounding. As a result, the SGS algorithms must depend quite sensitively on the data and even on the underlying distribution: tiny changes in the circumstances of the problem have big impacts on causal inferences.[8]

Exact conditional independence cannot be verified, even in principle, by statisticians using real data. Approximate conditional independence—which is knowable—has no consequences in the SGS scheme of things. That is one reason why the SGS theory is unrelated to the real problems of inference from limited data. The artificiality of the assumptions is the other reason.[9] For the moment, let us set these theoretical issues to the side, and look at the evidence cited for the success of the algorithms. SGS seem to offer abundant empirical proof for the efficacy of their methods: their book is studded with examples. However, the evidence is illusory. Many of the examples turn out to be simulations, where the computer generates the data. For instance, the ALARM network (p. 11 and pp. 145ff) is supposed to represent causal relations between variables relevant to hospital emergency rooms, and SGS claim (p. 11) to have discovered almost all of the adjacencies and edge directions 'from the sample data'. However, these 'sample data' are simulated; the hospitals and patients exist only in the computer program. The assumptions made by SGS are all satisfied by fiat, having been programmed into the computer: the question of whether they are satisfied in the real world is not addressed. After all, computer programs operate on numbers, not on blood pressures and pulmonary ventilation levels (two of the many evocative labels on nodes in the ALARM network). These kinds of simulations tell us very little about the extent to which modeling assumptions hold true for substantive applications. Moreover, arguments about causation seem out of place in the context of a computer program. What can it mean for one computer-generated variable to 'cause' another?

SGS use the health effects of smoking as a running example to illustrate their theory (pp. 18, 19, 75ff, 172ff, 179ff). However, that only creates another illusion. The causal diagrams are all hypothetical, no contact is made with data, and no substantive conclusions are drawn. Does smoking cause lung cancer and heart disease, among other illnesses? SGS appear not to believe the epidemiological evidence. They make their case using a rather old-fashioned method—a literature review with arguments in ordinary English (pp. 291–302). Causal models and search algorithms have disappeared. Thus, SGS elected not to use their analytical machinery on one of their leading examples. This is a remarkable omission. In the end, SGS do not make bottom-line judgments on the effects of smoking. Their principal conclusion is methodological: nobody besides them understood the issues.

> Neither side understood what uncontrolled studies could and could not determine about causal relations and the effects of interventions. The statisticians pretended to an understanding of causality and correlation they did not have; the epidemiologists resorted to informal and often irrelevant criteria, appeals to plausibility, and in the worst case to *ad hominem*. . . . While the statisticians didn't get the connection between causality and probability right, the. . . . 'epidemiological criteria for causality' were an intellectual disgrace, and the level of argument. . . was sometimes more worthy of literary critics than scientists. [pp. 301–2]

In this passage among others, scorn is heaped on investigators who have discovered important causal relations, like the health effects of smoking. The attitudes struck by SGS are quite extraordinary.

On pp. 132–52 and 243–250, SGS use their algorithms to analyze a number of real examples, mainly drawn from the social-science literature. What are the scoring rules? Apparently, SGS count a win if the algorithms more or less reproduce the original findings (rule #1); but they also count a win if their algorithms yield different findings (rule #2). This sort of empirical test is not particularly harsh.[10] Even so, the SGS algorithms reproduce original findings only if one is very selective in reading the computer output. We ran TETRAD on the four most solid-looking examples in SGS. The results were similar, and we report on one example here.[11] Rindfuss et al. (1980) developed a model to explain the process by which a woman decides how much education to get, and when to have her first child. The variables in the model are defined in Table 1.

Table 1. Variables in the model.[12]

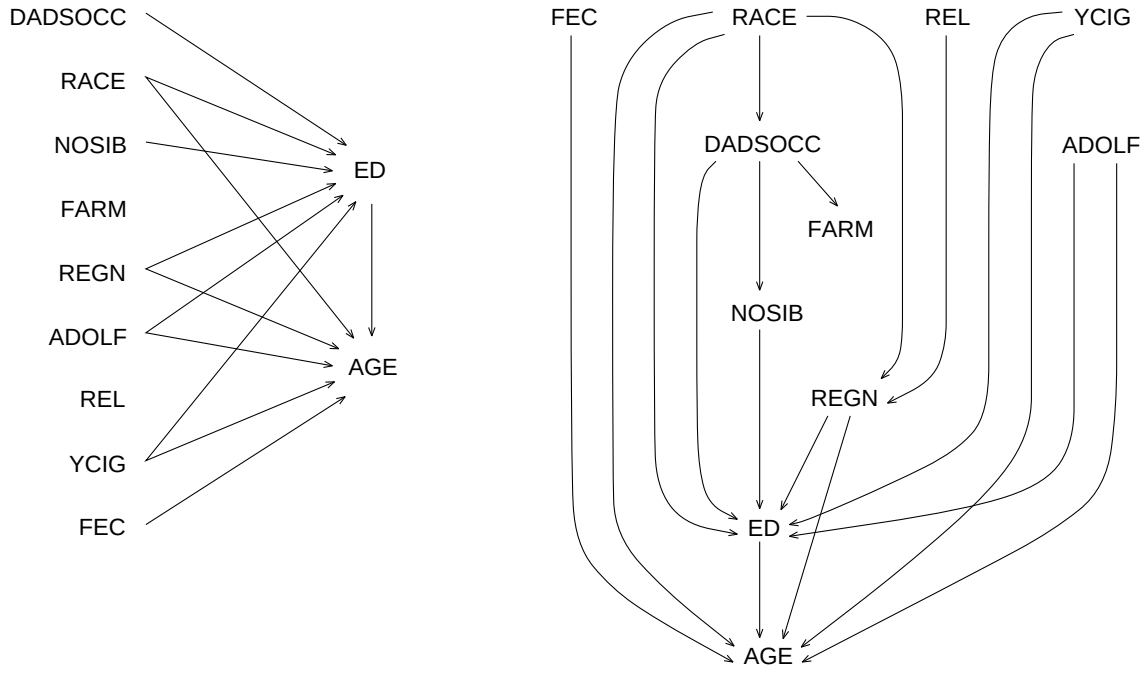
ED	Respondent's education (Years of schooling completed at first marriage)
AGE	Respondent's age at first birth
DADSOCC	Respondent's father's occupation
RACE	Race of respondent (Black=1, other=0)
NOSIB	Respondent's number of siblings
FARM	Farm background (coded 1 if respondent grew up on a farm, else 0)
REGN	Region where respondent grew up (South=1, other=0)
ADOLF	Broken family (coded 0 if both parents present at age 14, else 1)
REL	Religion (Catholic=1, other=0)
YCIG	Smoking (coded 1 if respondent smoked before age 16, else coded 0)
FEC	Fecundability (coded 1 if respondent had a miscarriage before first birth; else coded 0)

The statistical assumptions made by Rindfuss et al., let alone the stronger conditions used by SGS, may seem rather implausible if examined at all closely.[13] Here, we focus on the results of the data analysis. SGS report only a graphical version of their model:

> Given the prior information that ED and AGE are not causes of the other variables, the PC algorithm (using the .05 significance level for tests) directly finds the model [in the left hand panel of Figure 3] where connections among the regressors are not pictured. [p. 139]

Apparently, the left hand panel in Figure 3 is close to the model in Rindfuss et al., and SGS claim a victory under rule #1. However, this graph (published in *Causation, Prediction, and Search* p.140) is not the one actually produced by TETRAD II. The unedited graph is shown in the right hand panel of Figure 3. This graph says, for instance, that race and religion cause region of residence. Comments on the sociology may be unnecessary, but consider the arithmetic. REGN takes only two values (Table 1), so it cannot be a linear combination of prior variables with an additive Gaussian error, as required by TETRAD's statistical assumptions. FARM creates a similar problem. So does NOSIB. In short, the SGS algorithms have produced a model that fails the most basic test—internal consistency. Even by the relaxed standards of the social science literature, Figure 3 is a minor disaster.

Figure 3. The left hand panel shows the model reported by SGS (p. 140). The right hand panel shows the whole graph produced by the SGS search program TETRAD II.[14]



SGS seem to endorse the Automation Principle: the only worthwhile knowledge is the knowledge that can be taught to a computer. This principle is perverse. Despite SGS's agnosticism, the epidemiologists discovered an important truth—smoking is bad for you.[15] The epidemiologists made this discovery by looking at the data and using their brains, two skills that are not readily automated. SGS, on the other hand, taught their computer to discover Figure 3. The examples in SGS count against the Automation Principle, not for it.

2. Korb and Wallace

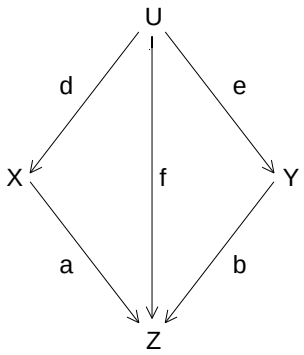
Most of the arguments in section 1 were presented in Humphreys and Freedman (1996); there were responses by Korb and Wallace (1997) and by Spirtes, Glymour, and Scheines (1997). We consider these responses in turn, replying only to a reasonable cross-section of the arguments. Although they seek to defend SGS, Korb and Wallace (1997) agree with us in many respects. The effort by researchers in artificial intelligence to automate the causal discovery process "promises in its most glorious moments to be as revolutionary to society *as might have been* a philosopher's stone" (p. 543, emphasis supplied); "the idea of automating induction perhaps appears merely foolish", "the exaggerations of Herbert Simon and Allen Newell ... are notorious" (p. 544), and

> The search for the philosopher's stone of an algorithm for induction may be a kind of perversion, but that is no more reason to end it now than the like charge would have been to prematurely end alchemy or Aristotelian physics when they were young. The search for a new science of induction is a glorious perversion. [p. 551–2]

2.1 Small correlations and instability

There are some minor points of disagreement. For instance, we say that the SGS algorithms depend quite sensitively on the data and the underlying distributions; a correlation of 0.000 precludes certain kinds of confounding while a correlation of 0.001 has no such consequences. Korb and Wallace (p. 549) respond that "the significance of weak [correlations] depends upon the size of the sample .... For small samples correlations equal to 0.001 have no implications for causal inference as they will not be detectable; for sufficiently large samples .... they will be unavoidably obvious". This is a technical misunderstanding: (i) to make a correlation of 0.001 "unavoidably obvious", sample sizes in excess of 1,000,000 would be needed;[16] (ii) the difficulty remains even if population data are available, so that estimation is unnecessary because the joint distribution is known. Point (ii) is somewhat technical: to explain it, we have to outline the SGS scheme for handling unmeasured confounders.

Figure 4. Variables X, Y, Z are observable; U is not observed; arrows represent causation; the lower-case letters next to the arrows quantify causal effects.



Consider, for instance, the path diagram in Figure 4.[17] Here, X, Y, Z are measured; U is an unmeasured 'confounder'. The parameter of interest is b, the causal effect of Y on Z. Suppose—as is the case most favorable to the SGS algorithms—that the joint distribution of (X, Y, Z) is known without error, i.e., population-level correlations can be determined. On the other hand, the full joint distribution of (X, Y, Z, U) is not known, corresponding to the idea that U is unmeasured. Ordinarily, b cannot be determined from the joint distribution of (X, Y, Z) because the influences of U and Y on Z cannot be separated. Suppose, however, that X and Z are conditionally independent given Y: the correlation between X and Z given Y is exactly 0, at the population level. Suppose also that the joint distribution of (X, Y, Z, U) is 'faithful' to the graph in Figure 4. Then b can indeed be computed, by regressing Z on Y. On the other hand, if the conditional correlation is .001 rather than .000 exactly, the regression coefficient can be biased to any arbitrary degree.[18]

For such reasons, the faithfulness assumption and exact conditional independence play a large role in the SGS theory. Of course, exact conditional independence cannot be determined from any finite sample. Instead, SGS test the 'null hypothesis' of independence, at some conventional significance level like .05. Unless the null hypothesis can be rejected, SGS adopt it—although there is substantial probability of error here, depending on the size of the sample and the values of the various parameters in the model.[19] Korb and Wallace have not responded to the point: correlations of .000 and .001—at the population level—play very different roles in the SGS theory. A sample of realistic size cannot distinguish between such correlations.

2.2 Testing the algorithms: simulations

Computer simulations can reveal the operating characteristics (e.g., error probabilities) of statistical procedures, given assumptions about the process generating the data. But simulations can hardly reveal whether the assumptions hold for real data sets. Among other things, what can it mean for one computer variable to 'cause' another? Korb and Wallace make two responses (pp. 546–47): (i) assumptions like linearity and independence give a reasonable starting point, used by working social scientists; (ii) "there is certainly no difficulty understanding how the values of some variables within computer programs may stochastically affect the values of other variables, for that is an elementary matter of computer programming".

The first is a well-worn argument: you have to learn to walk before you can run. The reasoning would be stronger if the modelers could more easily be seen to be making real progress. Point (ii) is assertion not explanation, and the problem is by no means 'elementary'. Computer programs represent *arithmetic* relations between variables. Although the programs are (in principle) deterministic, they can simulate random quantities, so that independent draws can be generated from a normal distribution. Thus, one can generate (X, Y) to be jointly normal, by setting

$$Y = a + bX + \epsilon, \tag{1}$$

where X and ϵ are independent normal variables and the latter has mean 0. In equation (1), a and b are parameters—numerical constants to be chosen by the programmer. Korb

and Wallace might say that X 'stochastically affects' Y. But this is too hasty. It would be trivial to rewrite the program so that (X, Y) have the same joint distribution, but are generated as

$$X = c + dY + \delta, \tag{2}$$

δ being independent of Y and having mean 0. Now Y 'stochastically affects' X. In short, the program does not determine causation.[20] The big point is this. The ability of an algorithm to infer causation must be tested on real examples not simulations—because the issue, in the end, is the extent to which the assumptions behind the algorithm hold for real data. After all, computer simulations are entirely under the control of the programmers, and can be designed so that statistical assumptions hold true. The real world is not so plastic.[21]

2.3 Real examples

To show the limitations of the SGS (1993) algorithms, we discussed work of Rindfuss et al. (1980) on the process by which a woman decides how much education to get (ED) and when to have her first child (AGE). Other variables are defined in Table 1. Rindfuss et al. concluded that the sort of woman who drops out of school to have children would drop out anyway: the line of causation runs from ED to AGE, not AGE to ED. Two-stage least squares was used to estimate the model. SGS (1993) reanalyzed the data using their computerized search algorithm 'TETRAD II'. Figure 3 shows on the left the model reported by SGS (1993, p. 140), and on the right the model we found using TETRAD II.[22]

Among other difficulties, the arrows from RACE and REL to REGN make no sense on substantive grounds and are impossible on arithmetic grounds. Indeed, the graph implies the equation

$$\text{REGN} = a + b\text{RACE} + c\text{REL} + \epsilon, \tag{3}$$

where by prior assumption ϵ is normal with mean 0 and is independent of RACE and REL. This is self-contradictory, because the left side of equation (3) is 0 or 1, while the right side must take all real values.

Korb and Wallace (pp. 550–51) have two main objections to this line of argument: (i) we should have told TETRAD II not to permit the troublesome causal arrows, and (ii) equation (3) is fine: "The fact that the true causal relationship is not linear does not mean that it cannot be detected using tests that assume linearity". They conclude that Figure 3 is a success, showing that "TETRAD II can recover structural causal relationships". These arguments have a certain ingenuous charm. Causal discovery algorithms succeed when they are prevented from making mistakes. If the algorithms work, they work *despite* failures in assumptions—and if they do not work, that is *because of* failures in assumptions. Arithmetic impossibility is brushed aside, and silliness is taken for truth. SGS used TETRAD II on the data and declared victory; Korb and Wallace endorse the claim. But TETRAD II generated absurd conclusions. It recovered some causal relationships that may be true, and it also recovered some that are plainly false. This is not a reliable

algorithm; and the standards used by Korb and Wallace to evaluate such algorithms are far too elastic.

2.4 Unarticulated objections

Korb and Wallace characterize our objections to the causal discovery research program as unarticulated:

> ... we wish the objectors would advance their reasons openly.... if these objectors would share their reasons then they might be subjected to the open criticism which seems to be necessary to the advancement of human knowledge. In the meantime, we suggest that causal theories as they have been presented in the causal discovery literature (supposing that the limitations of the simplifying assumptions are later overcome) are fully satisfactory to do the work we can reasonably expect of causal knowledge—namely, understanding, manipulating, and predicting events in the physical world. (p. 551)

We applaud the test—"understanding, manipulating, and predicting events in the physical world". But the argument will not do. The parenthetical supposition begs the central question. The linear-models approach to social science and its limitations have been debated at least since the Keynes-Tinbergen exchange.[23] The assumptions in SGS (1993) only compound the difficulties.[24] Let Korb and Wallace respond to the material already published. Even better, let them bring to the table some real examples showing the success of causal discovery algorithms.

3. Spirtes, Glymour, and Scheines

SGS (1997) defend themselves mainly by asserting that our summary of their work (Humphreys and Freedman, 1996) is incomplete or unfair. To make the argument, however, they do violence both to our position and to their own. We provide a number of illustrations, but it would take too long to answer each and every charge. SGS also raise substantive issues, and we respond to the main ones.

3.1 Circular definitions

We say,

> SGS do not give a reductive definition of 'A causes B' in non-causal terms. And their axiomatics require that you already understand what causes are. Indeed, the Causal Markov condition and the faithfulness assumption boil down to this: direct causes can be represented by arrows when the data are faithful to the true causal graph that generates the data. In short, causation is defined in terms of causation. [Humphreys and Freedman, 1996, p. 116.]

SGS (1997, p. 558) respond to this passage as if it accused them of taking the causal Markov condition and the faithfulness assumption as part of the meaning of causation. But that is mistaken. We can try to make our point more clearly. In the SGS setup, the directed acyclic graphs and associated random variables are mathematical objects. The Markov condition and the faithfulness assumption simply constrain these objects. The

causal Markov condition, in contrast, requires a graph to be causal (SGS, 1993, p. 54). Thus, to understand what counts as a causal graph, you must already understand what it is to be a cause (SGS, 1993, pp. 43, 45 ,47).[25]

3.2 The Automation Principle

According to the Automation Principle, the only worthwhile knowledge is the knowledge that can be taught to a computer. SGS (1997) disavow this idea: "We have never advocated such a principle, it plays no role in any argument that we have ever made ... [p. 558, f.n. 4]" And they take us sternly to task for not providing citations. However, SGS (1993) do try to make the case that computerized algorithms are the preferred methods for model selection, with social science theory having little role to play in this endeavor. (See especially pp. 127, 133, 137-8, 242.) For instance:

> In the social sciences there is a great deal of talk about the importance of 'theory' in constructing causal explanations.... In many of these cases the necessity of theory is badly exaggerated. [p. 133].
>
> Assuming the right variables have been measured, there is a straightforward solution to these problems [of model selection and causal inference]: apply the PC, FCI, or other reliable algorithm, and appropriate theorems from the preceding chapters, to determine which X variables influence the outcome Y, which do not, and for which the question cannot be answered from the measurements.... *No extra theory is required.* [p. 242, emphasis supplied]

For more recent material along the same lines, see Glymour (1997, pp. 214-15, 246). One passage deserves special attention:

> ... the general case for automated search is overwhelming: anything the human researcher actually knows before considering an actual data sample can be elicited and told to a computer; anything further the researcher can infer from examining a data sample can be inferred by a computer, and the computer can otherwise avoid the inconsistencies and the biases of humans. [p. 246]

That is the automation principle which SGS claim never to have advocated.

3.3 Reporting some findings but not others

In Figure 3, some of the output from TETRAD II conforms to the findings in Rindfuss et al., but some of the output is nonsensical. The latter was not reported by SGS in their book. SGS (1997, p. 565) find it appropriate to look only at the part of the output confirming their theses: "the part of the model that we displayed (SGS, 1993, p. 140) passes all of these tests". Indeed, SGS assert that displaying the rest of the output is "irrelevant" and runs "against common sense and the advice we give both in the book and in the program manual" (p. 565). This position must be rejected. Suppose that a model selection algorithm produces the following pair of equations,

$$Y = aX + \delta, \tag{4a}$$

$$Z = bU + cV + \epsilon; \tag{4b}$$

and it is believed on substantive grounds that (4a) is correct while (4b) is incorrect. SGS want to report only (4a)—and count a success for the algorithm. That is bad science. The evidence, considered as a whole, shows you cannot depend on the algorithm to select equations that are correct. Sometimes the algorithm picks a winner, sometimes it does not.

In Humphreys and Freedman (1996), we ran the algorithm at the .05 significance level, following SGS (1993, p. 139). Now SGS (1997, p. 565) suggest that the algorithm be run at significance levels .01, .05, .1 and .15. Presumably, arrows found at all these levels are 'robust', while others are suspect. We have followed their advice. Many of the arrows reported by SGS (1993, p. 140) are suspect, including those from RACE, NOSIB, REGN, and YCIG to ED. By comparison, troublesome arrows from REL and RACE to REGN are robust (right hand panel of Figure 3). The failures may be more robust than the successes.

3.4 Which version of the program did we use?

SGS (1997, p. 559) note that their computer package is called "TETRAD II" not "TETRAD". On this point, we concede error.[26] They find some inconsistency between the right hand panel of Figure 3, which we computed using TETRAD II, and the output from "the commercial version of the program", although they are not quite sure why that is; they conjecture that we used "a beta test version of the program" (p. 564). Back in the old days, after their book came out, we asked SGS for a copy of TETRAD II. Peter Spirtes kindly mailed one to us, on 8 October 1993. (The book was published earlier in 1993.) The program is well-behaved in many ways; among other things, it writes an identifier at the top of every output file:

```
COPYRIGHT (C) 1992
by Peter Spirtes, Richard Scheines,
Christopher Meek, and Clark Glymour.
All Rights Reserved

TETRAD II Version 2.1
```

SGS (1997, p. 559) also ask which of their algorithms we used; generally, we have specified the program module that produced the output.[27]

3.5 Other examples

Smoking

As we point out, SGS (1993, pp. 291–302) appear not to accept the standard epidemiological view that smoking causes lung cancer, heart disease, and many other illnesses. SGS (1997, p. 566) now reply that

> we never said the *evidence* did not support the conclusions; we said the *arguments* did not support the conclusions [emphasis in original].

This is a very nice distinction; in context, too nice. What can it mean? Perhaps SGS reviewed the underlying literature, reanalyzed the data, and found compelling new arguments to demonstrate the effects of smoking on health? Not so. In a previous response

to this sort of criticism, Spirtes and Scheines (1997, p. 174) say "We did not apply the algorithms to smoking and lung cancer data because we happened not to possess any such data". If SGS write again on this topic, we would like them to answer two simple questions. Do they agree that cigarettes kill? If so, why?

Armed Forces Qualification Test (AFQT)

Army recruits take a battery of 10 'subtests', including MC (Mechanical Comprehension), AR (arithmetic reasoning), WK (for word knowledge), GS (General Science), etc. The AFQT is a composite; but which subtests go into the composite? SGS (1993, pp. 243–44) contend that (i) this problem cannot be solved by ordinary regression methods, and (ii) the problem can be solved by TETRAD II. The first claim is false.[28] Now SGS (1997, p. 563) renew the second claim, ignoring the difficulties pointed out by Freedman (1997, p. 134): (i) left to its own devices, TETRAD II concludes that AFQT is the common cause of all the subtests; (ii) if instructed not to make this particular mistake, TETRAD II finds a cycle in the graph:

$$\text{MC} \to \text{AR} \to \text{WK} \to \text{GS} \to \text{MC}$$

In short, according the SGS algorithms, Mechanical Comprehension causes itself.[29]

Spartina

Spartina is a salt tolerant marsh grass with two forms, tall and short. Biomass (BIO) is in part a measure of the relative prevalence of the tall form. There are observational studies and a greenhouse experiment relating BIO to other factors, such as PH (low PH is acid, 7 is neutral, high PH is alkaline) and concentrations of various metallic salts, including magnesium, potassium, and phosphorus. SGS (1993, pp. 244–8) have reanalyzed data from one of the observational studies, and count the results as a success for their methods. SGS (1997) now reiterate the claim:

> PH was the controlling cause of the Spartina grass biomass (which was partially confirmed by experiment). [p. 563]

In 1993 SGS were actually a bit more cautious about this example:

> ... the only variable that may directly influence biomass in this population is PH .... PH and only PH can be directly connected with BIO.... the definition of the population in this case is unclear, and must in any case be drawn quite narrowly. [SGS (1993), p. 347]

Indeed, most the data were collected at PH below 5 or above 7, so results "cannot be extrapolated through PH values that approach neutrality" (p. 348). We have run TETRAD II on the data, and there is good reason to be cautious. The program cannot orient the edge between PH and BIO. In other words, the program cannot tell whether PH causes BIO, or BIO causes PH. However, two 'robust' conclusions can be drawn from the computer output: sodium causes magnesium, and *Spartina* does not need phosphorus to survive.[30]

3.6 Clinical trials

SGS (1997, p. 558) assert that the causal Markov and faithfulness conditions are "very widely (if implicitly) assumed in statistical and experimental reasoning, for example, in the

design and interpretation of randomized clinical trials". This claim is often made by SGS. See, for instance, pp. 555, 562; also see Spirtes and Scheines (1997, p. 167) or Scheines (1997, p. 191). However, it is not at all clear what the claim means. We pursued this question at some length with Richard Scheines and Peter Spirtes. As best we can see, they have two arguments.

> (i) Suppose a clinical trial finds no significant difference between the treatment group and the control group; on average over the subjects in the study, treatment has no detectable effect. However, a stronger inference is wanted: there is no effect of treatment on subgroups of subjects (older men, women with higher blood pressures, etc.). To justify this stronger inference from a finding of no overall effect, the faithfulness condition could indeed be used. However, when the clinical trials literature draws inferences about subgroups, it does so not by making aggregate comparisons, and certainly not by appealing to the faithfulness assumption. Instead, there is disaggregation—a direct comparison between subgroup members who are in the treatment and in the control conditions.[31]
>
> (ii) Suppose a clinical trial finds an effect for subjects in the study, and investigators wish to bolster the case by analyzing observational data. Then, according to Spirtes and Scheines, the causal Markov and faithfulness assumptions would come into play. That is quite debatable—ordinary working epidemiologists base conclusions on the strength of the effect they get, or its statistical significance, and on explicit control of confounders, but not on the basis of conditional independence. In any case, this second argument is diversionary, because it shifts ground from experiments to observational studies.[32]

As shown by Fisher and Neyman, and explained in many textbooks, randomization provides a secure basis for causal inference and statistical testing. There is no need for the faithfulness assumption or the rest of the SGS analytical apparatus. SGS can scarcely be serious when they assert (p. 556) that much of what we say about their methods "would equally be objections" to randomized clinical trials.

3.7 Statistical issues

Consistency vs. oracles

Statisticians have a concept of 'consistency': as more and more data come in, a consistent statistical estimator will get closer and closer to truth. This is widely viewed as a threshold criterion. There is another notion of 'Fisher consistency': when the estimator is applied to population data, it returns the parameter of interest. That notion is rarely used. SGS (1993) write as if they have proved their algorithms to be consistent (see, e.g., pp. 103–4); however, their theorems suggest Fisher consistency rather than consistency. In other words, their theorems say nothing about behavior with large, finite samples. This point was discussed by Freedman (1997, pp. 144–45, 180–81). The response by Spirtes and Scheines (1997, p. 170) does not clarify the issue.

Echoes of that debate may be found here. The SGS theorems require that exact conditional independence be known for the population. For this reason, SGS (1997) want an "oracle" on pp. 560–61, although the oracle segues into "a large sample limit" and

then a test based on a finite sample—thereby evading the basic question. They do concede (p. 560) that the theorems will "not guarantee success on finite samples". That is progress, but we would like an even bigger concession: the theorems do not show a probability of success tending to 1 as more and more data come in.[33]

Assumptions behind tests

> We say,
>
> The SGS algorithms for inferring causal relations from data are embodied in a computer program called TETRAD.... The program takes as input the joint distribution of the variables, and it searches over DAGs. In real applications, of course, the full joint distribution is unknown, and must be estimated from sample data. In its present incarnation, TETRAD can handle only two kinds of sample data, governed by conventional and unrealistic textbook models: (i) independent, identically distributed multivariate Gaussian observations, or, (ii) independent, identically distributed multinomial observations. These assumptions are not emphasized in SGS, but appear in the computer documentation and the computer output. [Humphreys and Freedman, 1996, p. 116.]

SGS appear to deny this—but in fact they concede it, for the "fully automated parts of the program", i.e., the algorithms they are using on the real examples (Rindfuss et al., AFQT, *Spartina*, etc.). SGS also seem to deny our claim that TETRAD II handles only "independent identically distributed samples"; they must be referring to the passage quoted above, although they substitute *Causation, Prediction, and Search* for TETRAD—and comment that more general samples are discussed at various points in the book. However, they again make the key concession: "statistical tests for conditional independence" are "not developed" except in the cases we mention.[34] In short, our account of the assumptions still seems to be right.

3.8 Are causal relations linear?

SGS (1997, p. 560) make this claim: "The normal family is almost coextensive with linear models, and so it is very odd that Humphreys, who has claimed that *all* causal relations are linear (Humphreys, 1989), should make such an objection". No page reference is given. In fact, no page reference could be given, for such a claim was never made. What Humphreys (1989, pp. 30–31) actually said was this: "for these reasons, I shall focus only on the case of $h_i(X_i) = b_i X_i$ in order to recover the linear models, while recognizing that this is a special case of causal relations"; "violations of linearity do not themselves preclude quantitative causal attributions".

3.9 Conclusion

SGS (1993) substantially overstates its case. The mathematical development has some technical interest, and the algorithms could find a limited role as heuristic devices for empirical workers—here is an alternative model to consider, there is a possible contradiction to ponder. Any much larger role could lead to considerable mischief, while claims to have developed a rigorous engine for inferring causation from association would be premature at

best. The theorems require an infinite amount of data, the assumptions are problematic, and the examples are unconvincing. We cannot agree that SGS have materially advanced the understanding of causality.

Notes

1. For more details, see Freedman (1997) and Humphreys (1997). Section 1 is a close paraphrase of Humphreys and Freedman (1996).

2. In this review, we try to give the intuition not the rigor; mathematical definitions are only sketched.

3. Statistical models for causation, in the sense of effects of hypothetical interventions, go back to Neyman (1923); also see Scheffé (1936) or Hodges and Lehmann (1964, Sec. 9.4). These writers were considering experiments, but the same models can be used to analyze observational studies—on the assumption that nature has randomized subjects to treatment or control. See, for instance, Rubin (1974) or Holland (1986, 1988). There is an extension to time-dependent concomitants and treatment strategies with indirect effects in Robins (1986, 1987ab). For discussion from various perspectives, see Pearl (1995).

Path models go back to Wright (1921), and were used by Blau and Duncan (1967) to analyze social science data. Graphical models for nonlinear covariation were developed in Darroch, Lauritzen, and Speed (1980). DAGs, conditional independence, and the Markov property were discussed in Kiiveri and Speed (1982, 1986); also see Carlin, Speed, and Kiiveri (1984). The 'd-separation' theorem expresses a graphical condition for conditional independence. See Pearl (1986) or Geiger, Verma, and Pearl (1990). Lauritzen et al. (1990) gave an alternative condition, which is easier to use in some examples. The 'faithfulness assumption' was introduced by Pearl (1988), and used for causal inference by Rebane and Pearl (1987); also see Pearl and Verma (1991). The 'manipulability theorem' appeared as the 'g-computation algorithm' in Robins (1986, 1987ab); also see Robins (1993). For a review of graphical models, see Lauritzen (1996). For a discussion of inferred causation, see Greenland, Pearl, and Robins (1998).

4. For causal inference, it is not enough that the distribution be faithful to *some* graph; the distribution must be faithful to the true causal graph that generates the data, the latter being a somewhat informal idea in SGS's framework. See Freedman (1997, sec. 12.3).

5. SGS (1993) justify their lack of an explicit definition by noting that probability theory has made progress despite notorious difficulties of interpretation. However, lack of clarity in the foundations of statistics may be one source of difficulties in applying the techniques. For discussion, see *Sociological Methodology 1991.*

6. The *causal representation convention* says: "A directed graph $G = \langle \mathbf{V}, \mathbf{E} \rangle$ represents a causally sufficient structure C for a population of units when the vertices of G denote the variables in C, and there is a directed edge from A to B in G if and only if A is a direct cause of B relative to $\mathbf{V}$. [SGS, 1993, p.47, footnote omitted.]" Following the chain of definitions, we have that "A set $\mathbf{V}$ of variables is **causally sufficient** for a population if and only if in the population every common cause of any two or more variables in $\mathbf{V}$ is in $\mathbf{V}$ or has the same value for all units in the population. [p.45, footnote omitted.]"

What constitutes a direct cause? "C is a **direct cause** of A relative to $\mathbf{V}$ just in case C is a member of some set $\mathbf{C}$ included in $\mathbf{V}\backslash\{A\}$ such that (i) the events in $\mathbf{C}$ are causes of A, (ii) the events in $\mathbf{C}$, were they to occur, would cause A no matter whether the events in $\mathbf{V}\backslash(\{A\} \cup C)$ were or were not to occur, and (iii) no proper subset of $\mathbf{C}$ satisfies (i) and (ii). [p.43]" This is only intelligible if you already know what causation means.

7. The most interesting examples are based on the assumption of a multivariate Gaussian distribution, and we focus on those examples. The documentation for TETRAD II is Spirtes, Scheines, Glymour and Meek (1993); point 2 on p.71 gives the statistical assumptions, which also appear on the computer printout. The algorithms are discussed in SGS pp.112ff, 165ff, and 183ff: these include the 'PC' and 'FCI' algorithms used in TETRAD II.

8. Thus, a correlation that equals 0.000 precludes certain kinds of confounding and permits causal inference; a correlation that equals 0.001 has no such consequences. For examples and discussion, see Freedman (1997, sec. 12.1), which develops work by J. Robins. Also see Section 2.1 below, and note 33.

9. The statistical assumptions—i.e., conditions on the joint distribution—include the Markov property and faithfulness, as noted in the text. For the algorithms to work efficiently and give meaningful output, the graph must be sparse, i.e., relatively few pairs of nodes are joined by arrows. Observations are assumed independent and identically distributed; the common distribution is multivariate Gaussian or multinomial. There is the further, non-statistical, assumption that arrows represent direct causes. This non-statistical assumption may be the most problematic: see the Summer, 1987 issue of the *Journal of Educational Statistics,* or the Winter, 1995 issue of *Foundations of Science.*

10. SGS (1993) eventually do acknowledge some drawbacks to their rules: "With simulated data the examples illustrate the properties of the algorithms on samples of realistic sizes. In the empirical cases we often do not know whether an algorithm produces the truth. [pp.132-133]"

11. The other examples are AFQT, *Spartina,* Timberlake and Williams (1984). The first two are discussed below.

12. The data are from a probability sample of 1,766 women 35–44 years of age residing in the continental United States; the sample was restricted to ever-married women with at least one child. DADSOCC was measured on Duncan's scale, combining information on education and income; missing values were imputed at the overall mean. SGS (1993) give the wrong definitions for NOSIB and ADOLF; the covariance matrix they report has incorrect entries (p.139).

13. See Freedman (1997, pp. 124ff).

14. The right hand panel is computed using the BUILD module in TETRAD II. BUILD asks whether it should assume 'causal sufficiency'. Without this assumption (note 5), the program output is uninformative; therefore, we told BUILD to make the assumption. Apparently, that is what SGS did for the Rindfuss example. Also see Spirtes et al. (1993, pp. 13–15). Data are from Rindfuss et al. (1980), not SGS; with the SGS covariance matrix, FARM 'causes' REGN and YCIG 'causes' ADOLF.

15. See Cornfield et al. (1959), International Agency for Research on Cancer (1986), U. S. Department of Health and Human Services (1990).

16. With sample of size n from a bivariate normal distribution, the sample correlation coefficient has a standard error on the order of $1/\sqrt{n}$. In the social sciences, samples of size 1,000,000 are few and far between.

17. This diagram is a hypothetical, representing three regression equations:

$$X = dU + \delta,$$
$$Y = eU + \epsilon,$$
$$Z = aX + bY + fU + \eta.$$

The variables U, X, Y, Z are normal with mean 0 and variance 1; the 'disturbance terms' δ, ϵ, and η are also normal with mean 0, independent of each other and of U. These are all assumptions.

18. As discussed above, 'faithfulness' is yet another assumption, precluding algebraic relations among the parameters a, b, d, e. For details, see Freedman (1997, pp. 114–19, 138–42, 148–49); also see SGS (1993, p. 35), Scheines (1997, pp. 193–94) or Glymour (1997, p. 209). On the degrees of possible bias, see Freedman (1997, pp. 148–149).

19. The chance of 'Type I error'—rejecting the null hypothesis when the latter holds true—is controlled to the level of .05. The chance of 'Type II error'—accepting the null when the latter is false—is then uncontrolled. Great sensitivity to the data must result from this approach: see note 33 below for the technical argument.

20. The parameters c and d in (2) can be obtained by regression, not by solving (1) for X in terms of Y. The calculation is standard; for details, see Freedman (1997, pp. 138–39, 157–9). If there are more variables and the faithfulness condition is imposed, the argument is more complicated (Freedman, 1997, pp. 150–51).

21. Korb and Wallace object to testing algorithms on real data, where difficulties are—apparently—only to be expected: "Humphreys and Freedman neglect to note that this is the natural outcome of their disfavor for testing with simulated data [p. 550]". Also see pp. 547–8.

22. Humphreys and Freedman (1996, pp. 119–21); Freedman (1997, pp. 121–32).

23. Keynes (1939, 1940); Tinbergen (1940). Other illustrative citations include Liu (1950), Meehl (1954), Lucas (1976), Manski (1993), Pearl (1995), or Abbott (1997). Also see Humphreys (1989); Freedman, Rothenberg, and Sutch (1983), Daggett and Freedman (1985), Freedman (1985, 1987, 1991, 1995). One major difficulty is specification of functional form: should the equations be linear, log linear, or something else?. Behavior of error terms is another problem: should these be assumed independent with common variance, or different variances, or dependent, and if the latter, in what way? Identifying and measuring the relevant confounders is yet another issue.

24. Freedman (1997); Humphreys (1997); Woodward (1997); Robins and Wasserman (1996).

25. Indeed, SGS are not in the business of defining causes: in their own words, they "advocate no definition of causation" (SGS, 1993, p. 41; SGS, 1997, p. 558). On the causal

Markov condition, see SGS (1993, p. 54), Scheines (1997, p. 196), Glymour (1997, p. 206). The definitional issues have substantive correlates. For one thing, the SGS treatment of examples (like Figures 3 and 4 above) does take for granted the causal Markov condition and faithfulness. More generally, imposing the causal Markov condition means that SGS cannot infer causation from association; at best, they can make causal inferences from prior assumptions about causation, and certain observed patterns of association in the data. SGS (1997) are clearer about this—and the statistical assumptions—than SGS (1993); see, for instance, p. 559. We may also be coming closer to agreement with SGS on another matter. We have pressed them to be more specific about their idea of causation: see, e.g., Humphreys (1997). SGS seem to be shifting to the stance that theirs is a manipulability account of causation: see Scheines (1997, p. 185) or Glymour (1997, p. 201).

26. SGS are also correct in saying their algorithms use not the t-test but the z-test (SGS, 1997, pp. 559, 561 and SGS, 1993, p. 128). However, differences between "observed significance levels" (P-values) from the two tests are generally minute. Take the Rindfuss et al. data. The correlation between DADSOC and YCIG is -0.043 with $n = 1766$ sample points, so $t = -1.808$ and $P = .0707$ while $z = -1.807$ and $P = .0708$. The correlation between ED and AGE is 0.380, so $t = 17.25$ while $z = 16.80$; either way, P is 0 to many decimal places. In this context, P represents the chance of getting a sample correlation as extreme as or more extreme than the observed one, assuming the 'null hypothesis' that the population correlation is 0. The null hypothesis is rejected if P is below some predetermined level. Absolute values of t or z are used to measure size, and the normal curve is used to compute the chance.

27. For instance, note 16 in Humphreys and Freedman (1996) explains that the graph for the Rindfuss et al. data was computed using the BUILD module in TETRAD II, with the assumption of causal sufficiency. That corresponds to the PC algorithm as used by SGS (1993, pp. 139–40). On the discrepancies between our output and theirs, see the notes to Figure 4.10 in Freedman (1997). Finally, if SGS write again on this topic, we have a question for them: which version of TETRAD II did they use for the data analysis in *Causation, Prediction, and Search*?

28. Freedman (1997, p. 133); for further arguments, see Spirtes and Scheines (1997, p. 174) and Freedman (1997, p. 178).

29. Freedman (1997, p. 133)

30. Presumably, if *Spartina* cannot grow in a certain environment, BIO is 0. SGS (1993, p. 247) recommend using the PC algorithm with significance levels of .05, .10, .15, and .20. At these levels, there is an arrow from sodium to magnesium. At the .05 and .10 level, phosphorus is an isolated node in the graph; at .15 and .20, it is separated from BIO. *Spartina* is a tough plant, but not that tough. If it can grow without a source of phosphorus, then it does not need DNA or RNA. (These molecules are built around a 'backbone' consisting of sugars and phosphates.) At the .20 level, and that level only, PH causes BIO. SGS give no reasons to justify their choice of significance levels; at the .01 level, BIO is an isolated node in the graph: Spartina's ability to grow is unaffected by any external causes. The sample size in this study is $5 \times 9 = 45$ (SGS, 1993, p. 244); reliable inferences from the algorithms cannot be expected (SGS, 1997, p. 563); i.e., correct

inferences as well as incorrect ones must be largely the result of chance. Why then do SGS present this example as a success? For more discussion, see Spirtes and Scheines (1997, p. 173) and Freedman (1997, p. 179). We thank Jeff Fehmi for background information on *Spartina*.

31. Of course, experiments are sometimes interpreted via regression models. For instance, it may be assumed that the response is $aX+\epsilon$, where X is 1 if the subject is in treatment and 0 otherwise, while ϵ is an error term. This assumption does not follow from randomization; if it is granted, then $a = 0$ does entail a universal finding of no effect; but it is hard to see where the faithfulness assumption comes into play.

32. SGS (1997, p. 562) now have a new argument: the statistical tests used in clinical trials, like the tests used by SGS, are valid only asymptotically. For instance, the 'level' of a test—the chance of rejecting the null hypothesis when the latter is true—may be set to .05; but in many cases, the calculation is only approximate, the approximation getting better and better as the sample size increases. The z-test used by SGS does have this characteristic, like the t-test often used in experiments. On the other hand, in many experiments, exact 'non-parametric' tests can be used. And the large-sample problems in the SGS program are in principle quite different from those in clinical trials. For example, in a wide variety of cases, the estimators in clinical trials are consistent; such theorems are not available for the SGS procedures, as discussed below. On the contrast between the SGS approach and ordinary epidemiologic approach, see Robins and Wasserman (1996).

33. To get such a theorem, one would need to make a sequence of tests with level tending to 0 and power tending to 1, calibrated to sample size and complexity of model. There is a further technical difficulty, because the output of the algorithm cannot depend continuously on the data: presence or absence of an arrow is binary (1 or 0), and the only continuous maps of Euclidean space into $\{0, 1\}$ are trivial, because Euclidean space is connected. The comments on p. 566 of SGS (1997) are not responsive to the technical issues.

34. Points 4 and 6 on pp. 559–60 of SGS (1997).

References

Abbott, A.: 1997, 'Of Time and Space: The Contemporary Relevance of the Chicago School', *Social Forces* 75, 1149–82.

Blau, P. M. and Duncan, O. D.: 1967, *The American Occupational Structure,* Wiley, New York.

Carlin, J. B., Speed, T. P., and Kiiveri, H. T.: 1984), 'Recursive Causal Models', *Journal of the Australian Mathematical Society* Series A, 36, 30–52

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L.: 1959, 'Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions', *Journal of the National Cancer Institute* 22, 173–203.

Daggett, R. and Freedman, D.: 1985, 'Econometrics and the Law: A Case Study in the Proof of Antitrust Damages', in L. LeCam and R. Olshen (eds.), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer,* Wadsworth, Belmont, Calif., Vol. I, pp. 126–75.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P.: 1980, 'Markov Fields and Log-Linear Interaction Models for Contingency Tables', *Annals of Statistics* 8, 522–39.

Freedman, D.: 1985, 'Statistics and the Scientific Method', in W. M. Mason and S. E. Fienberg (eds.), *Cohort Analysis in Social Research: Beyond the Identification Problem.* Springer-Verlag, New York, pp. 343–90. (With discussion.)

Freedman, D.: 1987, 'As Others See Us: A Case Study in Path Analysis', *Journal of Educational Statistics* 12, number 2. (With discussion.)

Freedman, D.: 1991, 'Statistical Models and Shoe Leather', in P. Marsden (ed.), *Sociological Methodology 1991,* American Sociological Association, Washington, D.C. (With discussion.)

Freedman, D.: 1995, 'Some Issues in the Foundation of Statistics', *Foundations of Science* 1, 19–83. (With discussion.) Reprinted in B. van Fraasen (ed.), *Some Issues in the Foundation of Statistics,* Kluwer, Dordrecht.

Freedman, D.: 1997, 'From Association to Causation via Regression', In V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 113–82. (With discussion.)

Freedman, D., Rothenberg, T., and Sutch, R.: 1983, 'On Energy Policy Models', *Journal of Business and Economic Statistics* 1, 24–36. (With discussion.)

Geiger, D., Verma, T. and Pearl, J.: 1990, 'Identifying Independence in Bayesian Networks', *Networks* 2, 507–534. Wiley, New York.

Glymour, C.: 1997, 'A Review of Recent Work on the Foundations of Causal Inference', in V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 201–48.

Greenland, S., Pearl, J., and Robins, J.: 1998, 'Causal Diagrams for Epidemiologic Research', to appear, *Epidemiology.*

Holland, P.: 1986, 'Statistics and Causal Inference', *Journal of the American Statistical Association* 8, 945–60.

Holland, P.: 1988, 'Causal Inference, Path Analysis, and Recursive Structural Equations Models', in C. Clogg (ed.), *Sociological Methodology 1988,* American Sociological Association, Washington, D.C., pp. 449-484.

Hodges, J. L., Jr. and Lehmann, E.: 1964, *Basic Concepts of Probability and Statistics.* Holden-Day, San Francisco.

Humphreys, P.: 1989, *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*, Princeton University Press.

Humphreys, P.: 1997, 'A Critical Appraisal of Causal Discovery Algorithms', in V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 249–63.

Humphreys, P. and Freedman, D.: 1996, 'The Grand Leap', *British Journal of the Philosophy of Science* 47, 113–23.

International Agency for Research on Cancer: 1986, *Tobacco Smoking.* Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Vol. 38. IARC, Lyon, France.

Keynes, J. M.: 1939, 'Professor Tinbergen's Method', *The Economic Journal* 49, 558–70.

Keynes, J. M.: 1940, 'Comment on Tinbergen's Response', *The Economic Journal* 50, 154–56.

Kiiveri, H.T. and Speed, T. P.: 1982, 'Structural Analysis of Multivariate Data: A Review', in S. Leinhardt (ed.), *Sociological Methodology 1982,* Jossey Bass, San Francisco.

Kiiveri, H. T. and Speed, T. P.: 1986, 'Gaussian Markov Distributions Over Finite Graphs', *Annals of Statistics* 14, 138–150.

Korb, K. B. and Wallace, C. S.: 1997, 'In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery', *British Journal of the Philosophy of Science* 48, 543-53.

Lauritzen, S. L.: 1996, *Graphical Models,* Oxford University Press.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer H.-G.: 1990, 'Independence Properties of Directed Markov Fields', *Networks* 20, 491–505, Wiley, New York.

Liu, T. C.: 1960, 'Under-identification, Structural Estimation, and Forecasting', *Econometrica* 28, 855-65.

Lucas, R. E. Jr.: 1976, 'Econometric Policy Evaluation: A Critique', in K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, vol. 1 of the Carnegie-Rochester Conferences on Public Policy, supplementary series to the Journal of Monetary Economics, North-Holland, Amsterdam, pp. 19–64. (With discussion.)

Manski, C. F.: 1993, 'Identification Problems in the Social Sciences'. In P. V. Marsden (ed.), *Sociological Methodology 1993,* Basil Blackwell, Oxford, pp. 1–56.

Meehl, P.: 1978, 'Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology* 46, 806–34.

Neyman, J.: 1923, 'Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes', *Roczniki Nauk Rolniczki* 10, 1–51, in Polish; English translation by D. Dabrowska and T. Speed: 1991, *Statistical Science* 5, 463-80.

Pearl, J.: 1986, 'Fusion Propagation and Structuring in Belief Networks', *Artificial Intelligence* 29, 241–288.

Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufmann, San Mateo, Calif.

Pearl, J.: 1995, 'Causal Diagrams for Empirical Research', *Biometrika* 82, 669–710. (With discussion.)

Pearl, J. and Verma, T.: 1991, 'A Theory of Inferred Causation', in J. A. Allen, R. Fikes, and E. Sandewall (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference,* Morgan Kaufmann, San Mateo, Calif., pp. 441–52.

Rebane, G. and Pearl, J.: 1987, 'The Recovery of Causal Poly-trees from Statistical Data', Proceedings, AAAI Workshop on Uncertainty in AI, Seattle, WA, pp. 222–228. Also in L. N. Kanal, T. S. Levitt, and J. F. Lemmer (eds.): 1989, *Uncertainty in Artificial Intelligence 3* Elsevier Science Publishers, Amsterdam, pp. 175–182.

Rindfuss, R. R., Bumpass, L. and St. John, C.: 1980, 'Education and Fertility: Implications for the Roles Women Occupy', *American Sociological Review* 45, 431–47.

Robins, J. M.: 1986, 'A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect', *Mathematical Modelling* 7, 1393–1512.

Robins, J. M.: 1987a, 'A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods', *Journal of Chronic Diseases* 40, Supplement 2, 139S–161S.

Robins, J. M.: 1987b, 'Addendum to "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect"', *Comp. Math. Applic.* 14, 923–45.

Robins, J. M.: 1993, 'Analytic Methods for Estimating HIV Treatment and Cofactor Effects', in D. G. Ostrow and R. Kessler (eds.), *Methodological Issues of AIDS Mental Health Research,* Plenum, New York.

Robins, J. M. and Wasserman, L.: 1996, 'On the Impossibility of Inferring Causation from Association without Background Knowledge', technical report no. 649, Statistics Department, CMU.

Rubin, D.: 1974, 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology* 66, 688–701.

Scheffé, H.: 1936, 'Models in the Analysis of Variance', *Annals of Mathematical Statistics* 27.

Scheines, R.: 1997, 'An Introduction to Causal Inference', in V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 185–99.

Spirtes, P., Glymour, C., and Scheines, R.: 1993, *Causation, Prediction and Search*. Springer Lecture Notes in Statistics, no. 81, Springer-Verlag, New York.

Spirtes, P., Glymour, C. and Scheines, R.: 1997, 'Reply to Humphreys and Freedman's Review of *Causation, Prediction, and Search*', *British Journal of the Philosophy of Science* 48, 555–68.

Spirtes, P. and Scheines, R.: 1997, 'Reply to Freedman', in V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 163–76.

Spirtes, P., Scheines, R., Glymour, C. and Meek, C.: 1993, *TETRAD II. Documentation for Version 2.2.* Technical Report, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Penn.

Timberlake, M. and Williams, K.: 1984, 'Dependence, Political Exclusion and Government Repression: Some Cross National Evidence', *American Sociological Review* 49, 141–46.

Tinbergen, J.: 1940, 'Reply to Keynes', *The Economic Journal* 50, 141–54.

U. S. Department of Health and Human Services: 1990, *The Health Benefits of Smoking Cessation, a Report of the Surgeon General,* Washington, D.C.

Woodward, J.: 1997, 'Causal Models, Probabilities, and Invariance', in V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, pp. 265-315.

Wright, S.: 1921, 'Correlation and Causation', *Journal of Agricultural Research* 20, 557–85.