An Experiment in Causal Discovery Using a Pneumonia Database

Peter Spirtes
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Greg Cooper
Department of Medical Informatics
University of Pittsburgh
Pittsburgh, PA 15213

Abstract

We tested a causal discovery algorithm on a database of pneumonia patients. The output of the causal discovery algorithm is a list of statements “A causes B”, where A and B are variables in the database, and a score indicating the degree of confidence in the statement. We compared the output of the algorithm with the opinions of physicians about whether A caused B or not. We found that the doctors opinions were independent of the output of the algorithm. However, an examination of the output of results suggested a simple, well motivated modification of the algorithm which would bring the output of the algorithm into high agreement with the physicians opinions.

1 THE PROBLEM

To make rational public policy decisions related to medicine requires knowledge of causal relations among variables. For example, in determining how many lives would be saved by treating everyone with pneumonia in hospitals, it is not enough to simply look at the probability of survival given hospitalization versus the probability of survival given no hospitalization. The reason is that generally the sicker patients are sent to the hospital, so that the influence of the hospitalization on the death rate is confounded with possibly unmeasured variables influencing how sick someone is. However, in many cases, it is not feasible to perform randomized trials for both ethical and practical reasons. Often, if we are to infer causal relations, we must do so from background knowledge and observational data.

We can describe this problem of causal inference somewhat more formally in the following way. A directed acyclic graph (DAG) G with a set of vertices **V** can be used to represent causal relations between variables, where an edge from A to B in G means that A is a direct cause of B relative to **V**; under this interpretation we call the DAG a causal DAG. (We assume that a causal DAG does not contain pairs of variables in which one is defined in terms of the other, or both are defined in terms of some third variable; this assumption is discussed at greater length in section 7.) If we assign a prior probability to each causal DAG, and a prior probability to the parameters of each causal DAG (representing the strengths of the causal connections), then given a database of patient records, it is possible to calculate the posterior probability of each causal DAG. The ideal Bayesian method of searching for causal relations among variables would be to simply write out the posterior probability of each causal DAG; the posterior probability of some variable X causing another variable Y could then be derived from the posterior probabilities of the DAGs. In practice, however, if the number of variables in a database is large, then it is not computationally feasible to calculate the posterior probability of each causal DAG, due to the astronomically large number of such DAGs.

2 PARTIAL SOLUTION

The following theorem follows simply from Cooper (1997) and Spirtes et al. (1995). (A measured variable V is **exogenous** if there is no variable which is a direct cause of V. A variable is **exogenous in a causal DAG** if there is no arrow directed into it.) We assume that there is no causal relation between the sampling mechanism and the measured variables (i.e., there is no selection bias).

Theorem: With probability 1 in the large sample limit, if

- each causal DAG containing the variables <E,A,B> in which E is exogenous has a non-zero prior probability,

- the prior probability of the parameters of each DAG is absolutely continuous with the BDe metric (Heckerman, et al., 1994),

- E is exogenous, and

- E → A → B has the highest posterior probability among all DAGs containing the variables <E,A,B> in which E is exogenous,

then in the true causal DAG, A is an ancestor of B (i.e. A is a cause of B) and there are no latent causes (i.e., unmeasured confounders) of A and B.

The importance of this theorem is that it gives a sufficient (but not necessary) condition for A being a cause of B, even without evaluating huge numbers of DAGs, and even when it is not known whether or not there may be unmeasured confounders.

There is a simple heuristic which can be used to reduce the number of triples for which the posterior probabilities are calculated. In particular, if E → A → B has the highest posterior probability among all DAGs containing the variables <E,A,B> in which E is exogenous, then (1) E, A and B are all highly dependent, and (2) E is independent of A given B (see Spirtes, et al. 1993.) We calculate the posteriors of DAGs only for triples of variables with these two properties. A more precise statement of the algorithm is given in the next section.

3 INSTRUMENTAL VARIABLE (IV) ALGORITHM

The IV algorithm takes as input background knowledge about which variables are exogenous, and a database consisting of patient records. An exogenous variable is also called an *instrumental variable*. The algorithm outputs a list of causal conclusions of the form "A causes B". The algorithm consists of the following steps:

1. Select a subset of variables **E** that are known to be exogenous. In the case of the pneumonia data (see below), the exogenous variables we used were race, age, and gender.

2. For each vertex E in **E**, search for measured variables A and B such that A is highly dependent on E, B is highly dependent on A, and E is independent of B given A. In the case of the data, we defined "highly dependent" to mean that the p value of the g^2 statistic measuring the dependence of discrete variables was less than 0.01, and "E is independent of B given A" means that the p value of the g^2 statistic measuring the conditional dependence of E and B given A is greater than 0.5.

3. For each triple of vertices <E,A,B> selected in step 2, for each DAG G that can be constructed out of the triple in which E is exogenous, calculate the posterior probability of G. If no DAG has a higher posterior probability than the DAG E → A → B then output "A causes B."

We assume each DAG compatible with the exogeneity of E has an equal prior probability. For each DAG, the prior probability over the parameters we used is the BDe prior described in Heckerman et al. (1994). The BDe prior assumes:

- that the data are complete, and
- that for any distinct variables X_1 and X_2, the set of parameters associated with X_1 and its parents are independent of the set of parameters associated with X_2 and its parents, and
- that for any two DAGs in which X_1 has the same parents the distribution over the parameters associated with X_1 and its parents are the same, and
- in a complete graph, the prior distribution over the parameters associated with a variable and its parents are Dirichlet.

Assuming equal prior probabilities for each DAG, we calculate the score as the natural log of the ratio of $P(G_1 | D)$ and $P(G_2 | D)$ where G_1 and G_2 are the DAGs with the highest and second highest posteriors, respectively. (In Table 1, this is the number in the column labeled Score.) This gives a rough idea how much the data D supports the conclusion that A causes B; the ratio between the highest and second highest posterior is generally large enough that this is a good approximation to carrying out the full calculation of the posterior.

4 DATA

The IV algorithm was tested on a pneumonia database of community acquired pneumonia patients (see Fine 1997 for details), which is called the pneumonia PORT database. Based on chart review, hundreds of data items were collected for each of the 2287 patients in the database.

A large number of variables had some missing values. A number of variables that had missing values were filled in with "normal" values. Even after this filling in, however, a number of other variables still had missing values. We selected a subset of 107 of the PORT variables for which a significant proportion of the population (1317 out of 2287 total) had no missing values for any variable in the subset. Step 2 of the IV algorithm was run on the subset of 107 variables for which the 1317 patients had complete records. However, for step 3 we did *not* use the subpopulation of 1317 that had no missing values. Rather, for each triple of variables chosen in step 3, the posterior of each causal DAG was calculated on the subpopulation of the patients who had no missing values for any variable *in that triple*; thus for particular triples, the sample size was slightly different, because there were different members of the population had missing values for different variables.

5 RESULTS

The IV algorithm was applied to the PORT database. The results obtained are shown in Table 1, listed in decreasing order of their scores (see Section 3). One pair was removed from the suggested list for reasons explained in section 7.

Table 1

Instrument	Cause	Effect	Score
age	coronary artery disease	myocardial infarction	18.41
age	current employment status	intravenous drug use (non-prescribed)	14.52
age	nausea	vomiting	9.28
gender	# of comorbid conditions	dire outcome (i.e., mortality or serious complications	8.47
gender	sputum	cough	7.99
age	current employment status	chronic obstructive pulmonary disease	7.55
age	current employment status	prior hospitalization within 30 days	4.87
age	current employment status	a history of chronic obstructive pulmonary disease requiring prior ICU admission	4.42
age	current employment status	days since last hospital discharge	0.56

6 PRELIMINARY ANALYSIS

As a preliminary test of the program's output, we asked a practicing physician at the University of Pittsburgh who sees pneumonia patients in his practice (Dr. Richard Ambrosino) to evaluate the output of the IV algorithm. Dr. Ambrosino had a close working knowledge of the pneumonia PORT database variables used in this study, because he had done prior (non-causal) research with this data. He was not, however, familiar with the IV algorithm. We presented this physician judge with a set of pairs of variables, some output by the algorithm as bearing a cause-effect relation to each other, and some chosen at random; the order of the pairs of variables was listed randomly. We asked the physician to classify each pair of variables into one of three classes: "Confident that A does cause B", "Don't know whether A causes B", or "Confident that A does not cause B." The results were that for all 10 pairs of variables suggested by the IV algorithm, the physician judge was confident that the relationship was cause and effect. For the randomly chosen pairs of variables, he was confident that the relationship between 5 of the 22 pairs was cause and effect; he was confident that 10 were not cause and effect; and in 7 he was not sure. Given this distribution of causal relations among the random pairs, Fisher's exact test of the independence of being chosen by the algorithm and being judged to be causal had a p-value of .0002 and can be strongly rejected.

In order to eliminate the hypothesis that the physician judge was simply taking all highly correlated pairs of variables as cause and effect, we submitted for his evaluation 15 more pairs of variables that were randomly selected from pairs of highly correlated variables (i.e. the g^2 statistic had a p-value of less than 0.01.) For these randomly chosen pairs of variables, the physician judge was confident that the relationship between 9 of 15 pairs was cause and effect; he was confident that 4 were not cause and effect; and 2 he was not sure of. Given this distribution of causal relations among the random pairs, if one chose 10 pairs of variables at random, an exact test of the independence of being chosen by the algorithm and being judged to be causal had a p-value of .0827, which is marginally significant.

7 ANALYSIS

The basic question we attempted to answer was: "Is the probability of A causing B given that the program says that A causes B higher than the probability of A causing B given that the program does not say that A causes B?"

However, there is a possible confounding factor that has to be considered. A necessary (but not sufficient) condition for the program to choose a pairs of variables is that they are highly associated (each pair passes a statistical test for association.) It is possible that the probability of A causing B among highly associated pairs of variables is much higher than the probability of A causing B among a random selection of pairs of variables. A second question that we attempted to answer was "Among highly associated pairs of variables, is the probability of A causing B given that the program says that A causes B higher than the probability of A causing B given that the program does not say that A causes B?"

One problem we faced was what to do with pairs of variables that are logically, rather than causally related. For example, the number of comorbid conditions is *defined* as the disjunction of cancer, swallowing difficulties, heart disease, etc. Sometimes two variables are both *defined* in terms of a third variable; e.g. agepresb and agepres6 are two different discretizations of age. When variables are logically related, there is generally a correlation between them, even though they are not causally related. The IV algorithm does not distinguish between logically related and causally related variables. In general, we assume that it is easy to find out whether two variables are logically related, so we do not count such pairs as either a success or an error. One of the pairs the program output was swalldia and cnumcomo. Cnumcomo is defined as the disjunction of a number of conditions including swalldia, so we eliminated it from consideration.

In order to answer the main question, we chose a number of highly associated variable pairs that had not been selected by the program to be compared to the pairs of variables that were selected by the program. When the algorithm measures association between a pair of variables A and B it uses the p-value of the g^2 statistic. Under the assumption of independence, the g^2 statistic is defined as the sum over all cells of the observed value in each cell times the natural log of the ratio of the observed value in the cell to the expected value in the cell. Asymptotically, the g^2 statistic is distributed as a χ^2 distribution. However, we did not use the p-value of the g^2 statistic when selecting variable pairs not chosen by the algorithm. When two associations are both large, even if the difference between the two associations is also large, the differences in the p-values of the two associations may be extremely small (i.e. the two p-values would both be zero to many decimal places.) For that reason, in judging the association between A and B, instead of using p-values, we used a standard adjustment of the g^2 statistic. The adjustment divides the g^2 statistic by the product of the sample size, and the minimum of the number of categories in A minus 1 and the number of categories in B minus 1. (The sample size differed somewhat between variable pairs, because in computing the association between A and B, we used the subpopulation that had no missing values for A and B. Since the subpopulations used varied with the variables, they had slightly different sample sizes.)

We selected variable pairs to compare with the variable pairs selected by the algorithm in two different ways. First we selected the 9 variables pairs with the highest adjusted g^2 measure of association, that were not logically related, and that had not been selected by the algorithm. Second, we attempted to match each of the 9 variable pairs A and B selected by the algorithm with a random variable pair that was not selected by the program, whose variables were not logically related to each other, and that had the same adjusted g^2 measure of association to three decimal places as A and B. However, it turned out that the two highest adjusted g^2 measure of association for the pairs of variables selected by the program could not be matched in this way, because there were no variable pairs fitting the description. Instead for two highest adjusted g^2 measure of association for the pairs of variables selected by the program we simply chose variable pairs that matched the adjusted g^2 measure of association as closely as possible. When this was done, it turned out that three of the variable pairs selected by the first method were the same as three of the variable pairs selected by the second method. So overall, there were 15 pairs of variables that we selected to contrast with the variable pairs selected by the algorithm.

A second problem to be faced is that we do not have a "gold standard" for when A causes B. We decided to use physicians opinions about the causal relations as our "gold standard." We enlisted the help five faculty physicians who practice internal medicine at the University of Pittsburgh and/or the Oakland VA Hospital (in Pittsburgh, PA) and who see pneumonia patients in their practices. These physicians were given a list of pairs of variables A and B, and were asked to assess whether in their opinion A causes B (encoded as 1), A does not cause B (encoded as 3), or they do not know whether or not A causes B (encoded as 2). If in their opinion A causes B, they were asked whether in their opinion there is also a common cause of A and B, no common cause of A and B, or they do not know whether there is a common cause of A and B. The exact formulation of the question, and the instruction to the physicians is given in the Appendix. In order to see whether the physicians as a group were reliable, we performed the following score of inter-rater reliability (Fleiss, 1981).

Let k be the number of categories into which ratings are made (in this case $k = 3$.) Let m be the number of raters (5) and n be the sample size (24). p_j is the proportion of ratings in category j, and q_j is $1 - p_j$. x_{ij} is the number of ratings on subject i in category j. In that case

$$\kappa = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} p_j q_j}$$

and the standard error is

$$\frac{\sqrt{2}}{\sum_{j=1}^{k} p_j q_j \sqrt{nm(m-1)}} \times \sqrt{\left(\sum_{j=1}^{k} p_j q_j\right)^2 - \sum_{j=1}^{k} p_j q_j (q_j - p_j)}$$

If there is no subject-to-subject variation in the proportion of positive ratings (and the proportion is not 0 or 1) then there is more disagreement within subjects than between subjects, and κ assumes the minimum value of $-1/(m\text{-}1)$. κ assumes the value 0 when the observed rate of agreement is that expected from chance. κ assumes the value 1 when there is perfect agreement among the raters. In this case, κ =.352 and the standard error is .0461. Hence the ratio of κ and the standard error is 7.64 and the hypothesis that $\kappa = 0$ can be strongly rejected. However, while a $\kappa > .75$ is considered excellent agreement, and $\kappa > .40$ represents good agreement, κ =.352 is generally considered poor agreement. However, it should be pointed out that when one rater assigns "cause" and a second rater assigns "don't know", this is in some sense not an actual disagreement.

We pooled the physicians opinions in three different ways. One variable that represented the physicians pooled opinions was *Sum*: for each question this was the sum of the values recorded by the physicians. A second variable was *Vote*, which was calculated in two steps: The first step removed all of the "don't know" answers, and in the second step, if a majority of the opinions left were "causal", the vote was "causal", if a majority of the opinions left were "not causal", the vote was "not causal"; and otherwise the vote was "don't know". The third variable was *Agree*, which is more conservative than Vote: Agree is 1 if Sum is less than 7, Agree is 3 if Sum is greater than 13, and otherwise Agree is 2. Table 2 shows the results (where IV algorithm = 0 if and only if the pair was not selected by the IV algorithm.)

Table 2

	IV algorithm	
Vote	0	1
1	9	4
2	0	2
3	6	3

	IV algorithm	
Agree	0	1
1	5	3
2	6	4
3	4	2

These tables indicate that "not causal" occurs a smaller percentage of the time among the pairs suggested by the IV algorithm than among the pairs not suggested by the IV algorithm. However, a chi-squared statistical test of the hypothesis that being selected by the IV algorithm is independent of Vote has a p-value of .197, and a test of the hypothesis that being selected by the IV algorithm is independent of Agree has a p-value of .965. Fisher's exact test of the two independencies yields similar results. Neither of these is significant.

8 IMPROVING THE IV ALGORITHM

We used these results to improve the IV algorithm by changing it so that it does not select those ordered variable pairs that the physicians were most dubious were causal. Among the pairs selected by the IV algorithm, the pairs that the physicians were most dubious about are shown in Table 3.

There are a number of obviously relevant features that the more dubious pairs output by the IV algorithm have in common. (In the following the values of discrete variables are for convenience encoded as integers.)

- 4 of the 5 dubious causal relations have the 4 lowest scores.
- If the Bayes Information Criterion were used to score the models rather than the posterior probability, 2 of the dubious causal relations (the 2 with the lowest scores) would not have been suggested by the algorithm at all.
- All of the dubious effects contained categories with relatively few members: intravenous drug use (33 have value 1), days since last discharge (38 with value 0, 25 with value 1, 7 with value 2), chronic obstructive pulmonary disease (45 with value 1), prior hospitalization (136 with value 1), and chronic obstructive pulmonary disease intensive care unit admission (20 with value 1). This is in contrast with the effects chosen by the IV algorithm that the doctors agreed with: myocardial infarction (245 with value 1), vomiting (594 with value 1), prior cough (564 with value 1), and dire outcome (261 with value 1.) It is possible that these low counts either effect the statistical tests (as indicated in the next item) or that they are rare enough that doctors simply are not aware of actual causal relations.

Table 3

Cause	Effect	Vote	Agree	Sum
current employment status	intravenous drug use (non-prescribed)	2	2	10
current employment status	chronic obstructive pulmonary disease	3	3	15
current employment status	days since last discharge from hospital	2	2	10
current employment status	prior hospitalization within 30 days	3	3	12
current employment status	history of chronic obstructive pulmonary disease requiring prior admission to ICU	3	3	14

- When conducting statistical tests of the association of the cause with the effect, on four of the five dubious effects (intravenous drug use, days since last hospital discharge, chronic obstructive pulmonary disease, and a history of chronic obstructive pulmonary disease requiring prior ICU admission) S-Plus issued a warning that the chi-squared test of independence may not be appropriate because the expected value of some cells was less than 5. It did not issue this warning on any of the 4 non-dubious effects.

Another obvious feature that all of the dubious pairs have in common is that the cause is current employment status. However, examination of current employment status revealed nothing unusual about its distribution, other than it had 4 categories, which is more than most of the variables in the database. In combination with the low counts in some of the categories of the dubious "effects", this produces statistical problems in testing dependence of the "cause" and the "effect".

These features suggest that the performance of the IV algorithm could be improved by eliminating pairs of variables for which the test of independence is dubious because some expected cell sizes are less than 5, and/or by raising the score threshold of what is considered a positive result for the algorithm.

There is a tradeoff here in changing the output of the IV algorithm to output fewer variable pairs; this leads to less information being output. The algorithm is already outputting relatively few pairs of variables, and the suggested changes would output even fewer.

If the IV algorithm were modified in these ways, it would choose only pairs of variables that the physicians were confident were causally related.

Because we have suggested changes to the algorithm in response to an examination of the data, we cannot properly test the algorithm on this data set. We plan to test the modified algorithm on other data sets.

9 APPENDIX: THE BACKGROUND SECTION OF THE CAUSALITY ASSESSMENT FORM

We are investigating statistical methods that attempt to uncover causal relationships from medical data. As a preliminary evaluation of these methods, we would very much appreciate your judgments of the 24 pairs below. Some of the pairs were generated by the statistical method and some were obtained by other means, including random generation. The order in which a pair appears has been randomized, so that the order contains no information about how a pair was generated.

The following is an example of the format in which the pairs of variables are listed.

Example:

A. patient has a fever during hospital admission
B. patient dies within 30 days of admission
___ (1) Confident that A does causally influence B.
In this case please also indicate whether you are:
a. ___ Confident A and B also are being influenced by a common cause.
b. ___ Don't know whether A and B are being influenced by a common cause.
c. ___ Confident A and B are not being influenced by a common cause.
___ (2) Don't know whether A causally influences B or not.
___ (3) Confident that A does not causally influence B.

The first line in the example contains a purported causal influence, which is always labeled as variable A. The second line contains the purported effect, which is always labeled as variable B. The remaining lines contain your judgment about the actual relationship between the two variables; you should mark exactly one of the entries labeled (1), (2), or (3). If you mark (1), then please further mark exactly one of (a), (b), or (c).

Note that we would say that fever "causally influences" mortality, even if fever actually prevents death within 30 days; that is, we count a variable that either promotes or suppresses another variable as a causal influence. Furthermore, to say that fever causally influences death within 30 days does not mean that it alone causally influences death within 30 days; simply that possibly in conjunction with other conditions found in the population it causally influences death within 30 days.

In asking (for example) whether fever causally influences death within 30 days, we mean to ask roughly:

> If it were possible to do a randomized clinical trial in which the treatment group were assigned to have a fever induced, and the control group were assigned to have a normal temperature maintained, are you confident that the number of deaths within 30 days of admission would be significantly different in the two groups?

We realize that it may not be practical, clinically useful, or even ethical to experimentally test some of the 24 relationships in the list given below. We are asking you, however, to provide your best judgment about what the relationship would be found to be, if such experiments were done, hypothetically.

We also realize that the assessments we are requesting leave matters somewhat vague, such as your confidence that a relationship is causal and your estimate of the strength of any relationship judged to be causal. We believe that to specify things too exactly would make for lengthy and somewhat unintuitive assessments; we would like to initially obtain simple assessments of causal relations.

In forming your causal assessments, assume a population of patients in North America who have community acquired pneumonia and are being seen at initial presentation.

Acknowledgements

We thank the pneumonia PORT project and its Principal Investigator, Dr. Wishwa Kapoor, for access to the data used in this study. We benefited from discussions with Dr. Michael Fine about the design of the causality assessment form. For generously giving their time in completing the assessment form, we thank Drs. Richard Ambrosino, Rosemarie Conigliaro, Keith Kanel, David Macpherson, Mark Roberts, and Jeff Whittle.

The research presented here was supported in part by grants BES-9315428 from the National Science Foundation and by grant LM05291 from the National

Library of Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Library of Medicine.

References

G.F. Cooper, "A simple algorithm for efficiently mining observational databases for causal relationships", *Journal of Data Mining and Knowledge Discovery*, 1997.

M. Fine, T. Auble, D. Yealy, B. Hanusa, L. Weissfeld, D. Singer, C. Coley, T. Marrie, W. Kapoor, "A prediction rule to identify low-risk patients with community-acquired pneumonia", *New England Journal of Medicine*, 336, 243-250. 1997.

J. Fleiss, *Statistical Methods for Rates and Proportions*, Wiley & Sons, New York, 1981.

D. Heckerman, D. Geiger and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data", *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 293-302, Morgan Kaufmann, 1994.

P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search*, New York, N.Y., Springer-Verlag, 1993.

P. Spirtes, C. Meek and T. Richardson, "Causal inference in the presence of latent variables and selection bias", *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. by Philippe Besnard and Steve Hanks, Morgan Kaufmann Publishers, Inc., San Mateo, 1995.