Human Causal Discovery from Observational Data

Ahmad I. Hashem, M.D.[1], and Gregory F. Cooper, M.D., Ph.D.[2]
[1] Section of Medical Informatics & Learning Research and Development Center
[2] Section of Medical Informatics & Intelligent Systems Program
University of Pittsburgh, Pittsburgh, PA 15260

Utilizing Bayesian belief networks as a model of causality, we examined medical students' ability to discover causal relationships from observational data. Nine sets of patient cases were generated from relatively simple causal belief networks by stochastic simulation. Twenty participants examined the data sets and attempted to discover the underlying causal relationships. Performance was poor in general, except at discovering the absence of a causal relationship. This work supports the potential for combining human and computer methods for causal discovery.

INTRODUCTION

Given observational data about the presence and absence of a side effect after administering or not administering a drug in a population of patients, how do workers in the health fields decide whether the side effect is caused by the drug? How accurate is such unaided human causal inference? In this paper, we report an experimental study that begins to address such questions by examining medical students' causal inference from observational data.

While there is still no universally accepted account for what constitutes normative causal inference, the probabilistic account has been receiving more attention in recent decades [1, 2]. Roughly, this account says that a cause is that which alters one's probability of the effect. The contingency paradigm [3-7, 15] that psychologists have used to study human causal inference is based upon this account.

In recent years, a more developed account of probabilistic causal inference based on a causal interpretation of Bayesian belief networks has been emerging [e.g., 8-10]. Causal belief networks (CBNs) are directed acyclic graphs whose arcs denote direct causal influences. The arcs are parameterized with a conditional probability distribution [11]. Nodes with no incoming arcs are given a prior probability distribution over values. Because of the recency of this account, little psychological research has been done to examine human causal inference when using Bayesian belief networks as a model of causality [but see 12]. One objective of the present study is to begin to fill this gap. This may provide better understanding of clinicians' cognition and suggest ways to augment human judgment of causality with automated causal discovery methods.

RELATED WORK

A typical study in the contingency paradigm asks participants to determine whether a drug causes a side effect. The information usually given to participants to aid them in answering this question [4, 6] consists of four values, which for the drug and side effect example would be: (*a*) the number of people who took the drug and had the side effect, (*b*) the number of people who took the drug and did not have the side effect, (*c*) the number of people who did not take the drug and had the side effect, and (*d*) the number of people who did not take the drug and did not have the side effect. In some studies, ratios are given instead of absolute numbers. Different models have been proposed to relate the participants' causal judgment to some or all of *a, b, c* and *d*. For example, Allan [13] proposed $a/(a+b) - c/(c+d)$ as a psychological model that captures the degree to which a person will believe that the drug causes the side effect. A variant of this model, the probabilistic contrast model, was advocated by Cheng and Novick [3].

A key problem in contingency paradigm studies is that they examine *covariation,* which, as acknowledged by contingency paradigm researchers [14], is an insufficient criterion for causal inference. The study we are reporting attempts to correct this by using causal belief networks as a model of causality -- a model that arguably makes explicit the necessary and sufficient conditions of normative causal inference from observational data [16, 8, 10].

METHOD

Participants

Twenty 2nd- and 3rd-year medical students at the University of Pittsburgh School of Medicine were recruited for the experiment. They were compensated $12 each.

Design

Our working assumption in this study is that CBNs define a set of causal processes from which we generate data. We generate sets of patient cases from CBNs and then test if, and how, participants can discover the causal relationships among measured variables from examining the cases. For our purpose, a "case" is a set consisting of a value for every variable in the CBN.

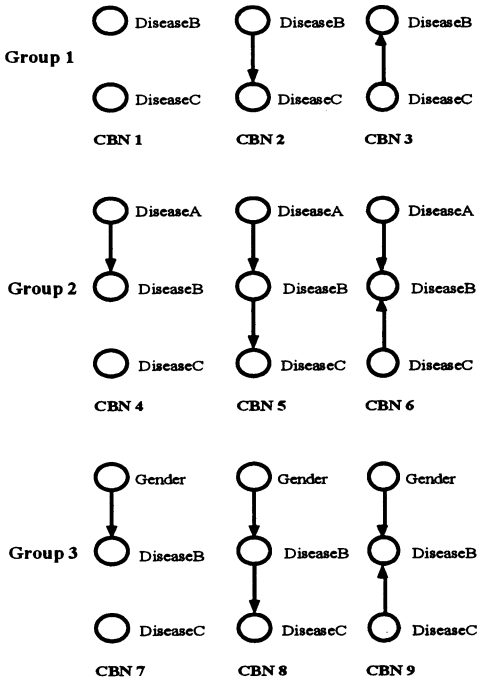

Figure 1. The nine causal belief networks used in the experiment.

We started by constructing nine relatively simple CBNs (see Figure 1), each containing either two or three binary nodes. In every CBN, the causal relationship of interest is that between *DiseaseB* and *DiseaseC*. The CBNs are divided into three groups, each consisting of three CBNs:

Group 1. CBNs consisting of two nodes only: *DiseaseB* and *DiseaseC*. The causal relationship of interest is from *DiseaseB* to *DiseaseC*, from *DiseaseC* to *DiseaseB*, or absent.

Group 2. CBNs consisting of three nodes: *DiseaseA*, *DiseaseB*, and *DiseaseC*. The only difference from Group 1 above is that *DiseaseA* causally influences *DiseaseB*. The participants were not provided with this information.

Group 3. CBNs consisting of three nodes: *Gender*, *DiseaseB*, and *DiseaseC*. The only difference from Group 1 above is that *Gender* causally influences *DiseaseB*. The participants were not provided with this information. They may, however, infer that while *Gender* may causally influence *DiseaseB* or *DiseaseC*, it cannot be causally influenced by either of them.

The probability distributions for the CBN nodes in Figure 1 are as follows:
- The prior probability of any node is 0.5.
- The posterior probabilities of a node x (such as *DiseaseB* in CBN5) that has one parent y are:

P (x=present | y=present) = 0.75
P (x=present | y=absent) = 0.25

- The posterior probabilities of a node x (such as *DiseaseB* in CBN6) that has two parents $y1$ and $y2$ are:

P (x | $y1$=present, and $y2$=present) = 0.8
P (x | $y1$=absent, and $y2$=present) = 0.6
P (x | $y1$=present, and $y2$=absent) = 0.4
P (x | $y1$=absent, and $y2$=absent) = 0.2

These probabilities define distributions with high variances; thus, the covariation among causally connected nodes is made more apparent, which may help the participants to discover causal relationships.

Using stochastic simulation [17], we generated nine unbiased data sets from the CBNs in Figure 1, each consisting of 1,000 cases. An example of a case generated for CBN7 is *Gender*=female, *DiseaseB*=absent, and *DiseaseC*=present. Participants were asked to view the nine data sets, one after the other, using a computer program designed for this experiment. The presentation order of the data sets was randomized, but data sets generated from Group 3

were presented last to avoid transferring the semantic information of *Gender* in Group 3 to *DiseaseA* in Group 2. Unique sequential integers were appended to the *Disease* variable names (replacing the letters *A*, *B*, and *C* in *DiseaseA*, *DiseaseB* and *DiseaseC*) to minimize anchoring and interference effects.

The screen of the computer program used in the experiment consists of 3 main sections:

The data manipulation section allows participants to view the frequencies for any variable's value given any (or no) value(s) for other variables in the data set. For a data set generated from Group 3, the participants may choose, for example, to view the probability of *DiseaseC*=absent given that *Gender*=male and *DiseaseB*=present.

The questions section prompts participants to enter their subjective degrees of belief that each of the following states holds: *DiseaseB* causally influences *DiseaseC*, *DiseaseC* causally influences *DiseaseB*, and *DiseaseB* and *DiseaseC* are not directly causally related. Each one of these three degrees of belief is a probability expressed as a percentage ranging from 0 to 100. A note on the screen reminds participants that the three entered numbers must add up to 100; the program does not accept the input if they do not.

The assumptions section informs participants that there are no cycles or feedback mechanisms underlying the relationships among the presented variables, and that the presented variables are not influenced by any "hidden" variables (i.e., unmeasured confounders).

Under the assumptions made in this study, it has been shown that causal relationships can be recovered from probabilistic dependencies [8-10, 18]. The gold standard (GS) we used to evaluate the participants' responses is the probabilities for causal relationships that a machine-learning algorithm [10] generates from the same sets of cases presented to participants. This GS is more appropriate for evaluation than the underlying CBN structure for two reasons: a) data sets generated stochastically may exhibit noise or sampling variation that complicate the process of inferring the underlying CBN structure; and b) some data sets are consistent with more than one underlying CBN structure. The GS provides the best possible answers given the data sets and the study assumptions. The number of variables in the data sets is small enough to justify exhaustive search over all possible structures (three for Group 1 data sets, 25 for Group 2, and 23 for Group 3 where there cannot be arcs from *DiseaseB* or *DiseaseC* to *Gender*). We assumed uniform priors over all structures. We used the likelihood equivalence scoring metric, assuming an equivalent sample size of one [18].

Procedure

Participants first receive a brief verbal description of the task and a demonstration on using the program. The program starts with an instruction screen followed by a practice session. The data sets are then displayed one after the other, with no possibility for viewing or modifying answers for previous sets. The program keeps track of all the activity the participants engage in to answer the questions. At the end of the experimental session, the program asks participants to enter comments on the strategy they used to answer the questions and on their definition of causality.

Material

The program used in this experiment runs on a Macintosh computer with 17" monitor. Participants were provided with a book of 9 blank, numbered sheets for optional use as scratch paper. The computer program prompts participants to turn to a new sheet at the end of every data set. Participants were instructed not to use a calculator.

RESULTS

From among the three answers the participants gave for every data set, we considered for analysis the answer to the question about the type of causal relationship between *DiseaseB* and *DiseaseC* that matched the structure of the CBN used in generating the data set. Thus, for CBN6 (see Figure 1), we considered for analysis the answers to the question about the degree of belief that *DiseaseC* causally influences *DiseaseB*. The percentages provided by participants were converted back to probabilities before analysis. Figure 2 shows the participants' mean answers along with the GS answers.

One hypothesis we examined is that the participants' answers are not different from the normative GS answers. A one-group t-test

showed that the difference between the observed and GS answers is significant at the 0.05 level for seven CBNs, and at the 0.06 level for eight. The participants' answers for CBN6 are not significantly different from GS.

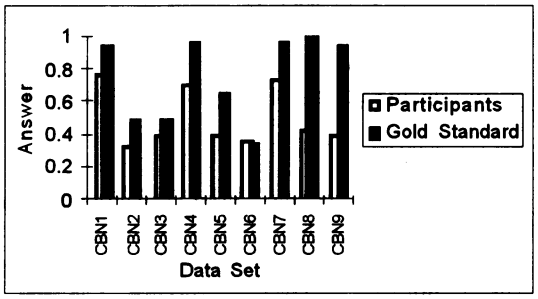

Figure 2. Mean answers of participants vs. GS.

Another hypothesis of interest is that the participants' answers are not different from answers generated by random guessing. We tested this hypothesis indirectly by first computing the expectation for the mean absolute difference between GS and a uniform random process assuming a uniform prior distribution between 0 and 1. This expectation, $GS^2 - GS + 1/2$, was then compared using a one-group t-test to the absolute difference between participants' answers and GS. The difference is significant at the 0.05 level for six CBNs, and at the 0.06 level for seven. It is not significant for CBN5 and CBN8. Figure 3 shows the differences from GS that were compared.

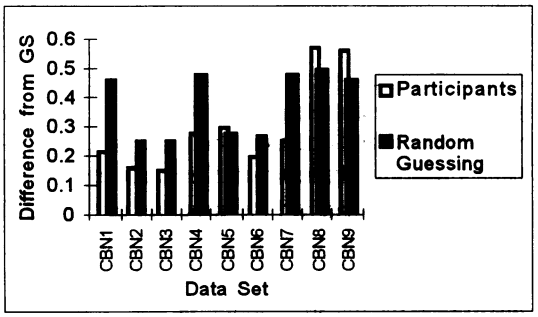

Figure 3. Mean absolute difference from GS: Participants vs. random guessing.

DISCUSSION

The results reported above support two conclusions. Participants' answers are generally significantly different from the normative answers and from random guessing. Further, the participants' answers are closer to GS than is random guessing when there is no causal relationship (CBN1, CBN4, CBN7) or when there are only two variables (CBN2, CBN3).

Note that, under our working assumptions, the direction of the causal relationship in CBN2 and CBN3 cannot be recovered from the data alone. A third variable is needed to untangle this direction [8]. The task in these networks is one of discovering statistical association more than causal relation. It can be said, then, that the participants' performance at discovering causal relationships is poor except at discovering the *absence* of such relationships (CBN1, CBN4, CBN7), which amounts to inferring the absence of a correlation.

We notice that the participants' answers for CBN6 are not significantly different from the normative answer. We attribute this to a sampling problem in generating the cases from CBN6 that led to poor performance by the machine-learning algorithm used as GS. We verified that the GS would be higher, thus possibly significantly different from the participants' answers, when using a different or larger sample of cases. This sampling problem gives further support to choosing GS for evaluating participants' answers.

We also notice that the participants' answers for CBN5, CBN8 and CBN9 are slightly worse (further from GS) than random guessing, significantly so in the case of CBN9. We attribute this to the participants' failure to utilize information about *DiseaseA* and *Gender* when answering questions concerning the causal relationship between *DiseaseB* and *DiseaseC*. In response to the question "How would you define the word causality?", the participants' answers were almost always formulated in terms of only two variables. Thus it seems that, unlike the machine-learning algorithm, participants answer questions about the causal relationship between two variables without considering information available about other variables.

In this exploratory study we did not control for testing of multiple hypotheses. In addition to such control in future studies, it would be of interest to do experiments using concrete real data of known drug side effects obtained from

randomized clinical trials. It would also be interesting to see if a different mode of presenting the data (such as sequential presentation of cases), some prior tutoring on CBNs, or studying a different population such as expert physicians, would influence the results. The current results seem to suggest that humans are not as adept at using only abstract data to infer causality while machines seem relatively good at this task. Humans' strengths may be more in expressing prior probabilities for likely causal relationships, based on the domain-specific *meaning* of the nodes. Investigating, developing, and testing methods for combining clinician and computer methods for causal discovery is an area that appears worth pursuing.

Acknowledgments

This research was funded in part by grant 5-T15-LM07059 from the National Library of Medicine and grant BES-9315428 from the National Science Foundation. We would like to thank Noetic Systems, Inc. for providing us with the Ergo™ software that was utilized in calculating probabilities.

References

1. Salmon, WC (1971). Statistical explanation. In W. Salmon (Ed.), *Statistical Explanation and Statistical Relevance.* Pittsburgh, PA: University of Pittsburgh Press.
2. Suppes, P (1970). *A Probabilistic Theory of Causality.* Amsterdam: North Holland.
3. Cheng, PW, & Novick, LR (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545-567.
4. Allan, LG, & Jenkins, HM (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology, 34,* 381-405.
5. Jenkins, HM, & Ward, WC (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79*(1), Whole No. 594.
6. Anderson, JR, & Sheu, CF (1995). Causal inference as perceptual judgments. Dept. of Psychology, Carnegie Mellon University.
7. Wasserman, EA, Dorner, WW, & Kao, SF (1990). Contribution of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 509-521.
8. Spirtes, P, Glymour, C, & Scheines, R (1993). *Causation, Prediction, and Search.* New York: Springer-Verlag.
9. Pearl, J, & Verma, T (1991). A theory of inferred causation. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Knowledge Representation and Reasoning: Proceedings of the Second International Conference,* 441-452. New York: Morgan Kaufmann.
10. Cooper, GF, & Herskovits, E (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9,* 309-347.
11. Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.
12. Ahn, WK, & Mooney, RJ (1995). Biases in refinement of existing causal knowledge. In J. D. Moore & J. F. Lehman (Eds.), *Seventeenth Annual Conference of the Cognitive Science Society,* 437-442. Hillsdale, NJ: Lawrence Erlbaum.
13. Allan, LG (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15,* 147-149.
14. Cheng, PW, & Novick, LR (1992). Covariation in natural causal induction. *Psychological Review, 99(2),* 365-382.
15. Allan, LG (1993). Human contingency judgment: Rule based or associative? *Psychological Bulletin, 114,* 435-448.
16. Heckerman, D (1995). A Bayesian approach to learning causal networks (Technical Report No. MSR-TR-95-04). Microsoft.
17. Henrion, M (1988). Propagating uncertainty in Bayesian networks by logic sampling. In J. Lemmer & L. Kanal (Eds.), *Uncertainty in Artificial Intelligence 2,* 149-163. Amsterdam: Elsevier.
18. Heckerman, D, Geiger, D, & Chickering, D (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* 293-301. San Mateo, CA: Morgan Kaufmann.