

# Causal Discovery from Data in the Presence of Selection Bias

Gregory F. Cooper  
Section of Medical Informatics  
University of Pittsburgh

## Abstract

Recent research advances have made it possible to consider using observational data to infer causal relationships among measured variables. Selection bias results from the observation of entities that are not representative of the entities that are generated by a causal process of interest. This paper shows that we can sometimes detect the presence of selection bias in observational data. The paper also demonstrates how selection bias can hinder the discovery of causal relationships from observational data. As we will describe, the use of experimental data (e.g., data from randomized, controlled trials) to discover causal relationships can be susceptible as well to problems involving selection bias. We offer suggestions for how to proceed with causal discovery in the face of selection bias.

## Introduction

The discovery and characterization of causal relationships among variables is a primary focus of much of science. Randomized, controlled experiments are the gold standard by which science establishes the presence of causal relationships. Such experiments, however, are often expensive and sometimes are not possible. In contrast, there is a relative bounty of observational data, which is growing rapidly in the form of computer-stored databases. Ideally, we would like to be able use observational data to infer causal relationships.

Researchers recently have developed methods for determining the presence or absence of causal relationships from statistical independence and dependence relationships extracted from observational data [16, 17, 20, 21, 22]. These elegant methods have the potential to provide scientists and others with powerful new tools for causal discovery. They complement experimental science by using observational data to (1) corroborate experimental results and (2) suggest new causal relationships that can be tested experimentally.

There are a number of known biases that can hinder the use of observational data for causal discovery [7, 19]. Selection bias is one such bias, and it is the focus of this paper. To define selection bias, let  $E_Q$  be a set of entities that are generated by a process  $Q$ . For example,  $E_Q$  might be a population of patients in the community. Let  $E_S$  be the set of sampled entities that we measure. More specifically, we measure particular properties of each member of  $E_S$ . Let  $Z$  be a set of variables that we use to express the properties of the entities in  $E_S$ . Assume we have a database containing the values of  $Z$  for each member of  $E_S$ . Continuing the example,  $E_S$  might be a set of patients who come to the emergency room of a community hospital, and  $Z$  might be a set of symptoms for which we gather information on each patient in order to create our database. We use  $P_S(Z)$  to denote the joint probability distribution over  $Z$  for the entities in  $E_S$ ;  $P_S(Z)$  is constructed from our database by applying some asymptotically convergent probability estimation procedure, which here we leave implicit. We will assume that process  $Q$  generates an entity with values  $\underline{Z}$  with probability  $P_Q(\underline{Z})$ , where  $\underline{Z}$  denotes an arbitrary instantiation of the variables in  $Z$ . Let  $P_Q(Z)$  be the joint probability distribution for process  $Q$ . We will say that our database on  $E_S$  exhibits *selection bias* if and only if  $P_S(Z) \neq P_Q(Z)$ .

As we have defined it, selection bias may exist due simply to sampling errors, particularly if the sample size (i.e.,  $|E_S|$ ) is small. This paper does not, however, focus on sample size as a source of selection bias. We focus instead on selection bias that persists even when the sample size becomes arbitrarily large. In medicine, selection bias has been termed *referral bias* by some researchers [7]. A well known version of referral bias is called Berkson's paradox (also known as Berkson's fallacy or Berkson's bias) [1]. Berkson's paradox and related forms of referral bias have been demonstrated empirically in several areas of medicine [3, 9, 13, 18].

We will first assume that the observational data  $D$  available to us consists of the values of the variables in  $Z$  from unbiased sampling of the entities in  $E_Q$ . Our goal is to use data  $D$  to discover causal relationships among the variables in  $Z$  for the entities in  $E_Q$ . Later we will consider causal



discovery from biased samples of  $E_Q$ . We also will describe how selection bias can interfere with causal discovery from experimental data.

## Causal Probabilistic Networks

In this section we briefly describe causal probabilistic networks. For a more formal introduction, see [15, 21]. Before introducing a causal probabilistic network, we first define a Bayesian belief network, which we call a Bayesian network for short. A *Bayesian network* is a directed acyclic graph  $B_{str}$  in which nodes represent model variables, and for each variable (node)  $x_i$  a probability distribution  $P(x_i | \pi_i)$  is specified, where  $\pi_i$  are the variables in the model corresponding to the parents of  $x_i$  in  $B_{str}$ . The representation of conditional dependence and independence among variables is the essential function of Bayesian networks. In particular, a Bayesian network  $B$  incorporates the following Markov assumption: For each variable  $x_i$  in  $B$ , if we condition on any set of values for the variables in  $\pi_i$ , then the probability distribution of  $x_i$  is independent of all the variables corresponding to non-descendants of node  $x_i$  in  $B_{str}$ . A node  $y$  is a descendant of  $x_i$  if there is a directed path from  $x_i$  to  $y$ .

A criterion called *d-separation* captures exactly the conditional independence and dependence relationships that follow from the Markov assumption above [8, 15]. The following definition of d-separation is taken from [16]. Let  $T, U$ , and  $V$  be disjoint subsets of the nodes in  $B_{str}$ . Let  $p$  be any path between a node in  $T$  and a node in  $U$ , where a path is any succession of arcs, regardless of their directions. We say a node  $w$  has converging arrows along a path if two arcs on the path point to  $w$ .  $V$  is said to block  $p$  if there is a node  $w$  on  $p$  satisfying one of the following two conditions: (1)  $w$  has converging arrows (along  $p$ ) and neither  $w$  nor any of its descendants are in  $V$ , or (2)  $w$  does not have converging arrows (along  $p$ ) and  $w$  is in  $V$ .  $V$  is said to d-separate  $T$  from  $U$  in  $B_{str}$  if and only if  $V$  blocks every path from a node in  $T$  to a node in  $U$ .

In this paper, we consider causal processes that can be modeled as Bayesian networks. A *causal probabilistic network* is a Bayesian network in which arcs in  $B_{str}$  denote direct causal dependence relative to the variables in  $Z$ . Informally, the notion that  $x$  directly causes  $y$  is as follows. There is an arc from  $x$  to  $y$  if and only if there is some manipulation of the value of  $x$  that would change our probability distribution over the value of  $y$ , conditioned on some state of the model variables other than  $x$  and  $y$ . For additional discussion of the notion of causal manipulation, see [16, 21].

## Discovery of Causal Probabilistic Networks from Observational Data

Suppose we are interested in knowing about a set of causal processes  $Q$  in nature. We will assume we can adequately model  $Q$  as a causal probabilistic network  $C$ . Let  $C_{str}$  denote the causal-network structure of  $C$ . The game we play is this. We assume that  $Q$  is equal to  $C$ . Our goal therefore is to reconstruct  $C_{str}$  from  $D$ .<sup>1</sup> From knowing  $C_{str}$ , we hope to gain some insight regarding  $Q$ . We denote the set of variables in  $C$  as  $Z$ . If  $T, U$ , and  $V$  are mutually exclusive subsets of the nodes in  $C$  and if  $V$  d-separates  $T$  from  $U$ , then we will write  $I_C(T, U | V)$ . If  $V$  is empty, we will write  $I_C(T, U)$ . If  $T, U$ , or  $V$  contains just a single variable, for simplicity we use that variable's name in  $I_C(\bullet)$ , rather than use set notation.

Recent research has investigated methods for deriving  $C_{str}$  from  $D$  based on statistical tests of independence and dependence [16, 17, 20, 21, 22]. More specifically, a statistical test  $t$  is applied to  $D$  in order to derive relationships of conditional dependence and independence among the variables in  $Z$ . For example, for discrete data,  $t$  might be a chi-square test with a particular, fixed significance level. We use  $I_{D,t}(T, U | V)$  to indicate that for test  $t$  applied to data  $D$ , each subset of the variables in  $T$  is independent of each subset of the variables in  $U$  given any values of the variables in  $V$ . The key assumption that allows us to derive (or partially derive)  $C_{str}$  from  $D$  is this:

For any  $T, U$ , and  $V$  that are mutually exclusive subsets of  $Z$ , it is the case that  $I_C(T, U | V)$  is true if and only if  $I_{D,t}(T, U | V)$  is true.

Furthermore, suppose the variables captured in  $D$  are only a subset  $Z'$  of the variables  $Z$  in  $C$ . That is,  $Z \setminus Z'$  are hidden variables in  $C$ . In this case, the key assumption above is modified by substituting  $Z'$  for  $Z$ , and we would likewise replace each corresponding instance of  $Z$  in this paper

<sup>1</sup> A related goal is to find the parameterization of  $C$ , although in this paper that is not our focus.



by  $Z'$ . For a detailed discussion of the assumptions underlying causal discovery from observational data, see [21].

The network structures in Table 1 represent the type of causal relationships which are the primary focus of this paper. While they are not exhaustive in scope, they do represent fundamental causal relationships. Note, we assume that (1) there is a known cause  $w$  of  $x$ , (2) a known cause  $z$  of  $y$ , and (3)  $w$  and  $z$  are connected (if at all) only by causal paths that go through both  $x$  and  $y$ . These causal relationships could be known *a priori* (based for example on randomized, controlled trials) or they could be learned from observational data (which would require measuring additional variables not shown). An important point worth emphasizing is that the causal discovery methods discussed here (that use observational data) depend on our considering a *web* of relationships among variables.

Table 1 contains the d-separation relationships in  $C_{str}$  that we assume correspond to the independence and dependence relationships found when applying test  $t$  to data  $D$ . A plus sign means the independence relationship holds, while a blank means it does not. Note that the set of relationships in each row is unique relative to the relationships in the other rows. Thus, we can distinguish among these four structures. We note that d-separation implies additional independence relationships not listed in Table 1; we will not need to consider them until later in this paper. For the causal-network structures in Table 1, we have the following theorem.

**Theorem 1.** The seven relationships listed in Table 1 are sufficient to distinguish the four causal structures shown there. Thus, if we assume we can infer these relationships by applying test  $t$  to data  $D$ , then observational data is sufficient to distinguish among these four causal structures.

**Proof** The distinguishing relationships in Table 1 were derived by applying the d-separation criterion to the respective structures shown. It is straightforward to verify the validity of these relationships. Since the pattern of the seven distinguishing relationships is unique for each of the four network structures in Table 1, these relationships are sufficient to distinguish among the structures.  $\square$

Note that the seven relationships listed in Table 1 are sufficient, but not altogether *necessary* to distinguish among the four network structures in that table. In particular,  $R_2$  and  $R_3$  are also sufficient; no smaller set of such binary relationships could distinguish among the four structures. We include in Table 1 all seven relationships because they will be useful in other sections of the paper.

Table 1. Variable  $h$  represents a hidden variable. In a given row, relationship  $R_i$  has a plus sign if, and only if, the network in that row exhibits the independence condition associated with  $R_i$ .

No.	Structure of a causal network $C$	Distinguishing relationships						
		$R_1$	$R_2$	$R_3$	$R_4$	$R_5$	$R_6$	$R_7$
1	$w \longrightarrow x \quad y \longleftarrow z$	+	+	+	+	+	+	+
2	$w \longrightarrow x \longrightarrow y \longleftarrow z$			+	+		+	
3	$w \longrightarrow x \longleftarrow y \longleftarrow z$		+			+		+
4	$w \longrightarrow x \begin{matrix} \swarrow h \\ \searrow \end{matrix} y \longleftarrow z$				+	+	+	+

Key:  $R_1: I_C(x, y)$ ,  $R_2: I_C(x, z | y)$ ,  $R_3: I_C(w, y | x)$ ,  $R_4: I_C(w, z | x)$ ,  $R_5: I_C(w, z | y)$ ,  $R_6: I_C(x, z)$ ,  $R_7: I_C(w, y)$

## Selection Bias in Causal Discovery

We have been assuming that  $D$  is generated from an unbiased stochastic sampling of points in the sample space defined by causal network  $C$ . In medicine, for example,  $D$  might represent a set of symptoms for each person in a random sampling of people in a particular community. We often may not have access to information about the unselected population  $E_Q$  as represented by  $D$ , but rather, access only to information about a subset  $E_S$  of  $E_Q$  that is selected by a process unknown to us; we will



use  $D_S$  to denote the information we have about the members of  $E_S$ . Thus,  $D_S$  is the data that we actually observe. In this paper,  $D_S$  contains data about the values of variables  $w, x, y$ , and  $z$ .  $D_S$  also contains a new variable  $s$ , which we introduce into  $Z$  in order to represent the sample selection process. If an entity  $e$  is in  $E_S$  then  $s$  has the value *selected* in the record in  $D_S$  that represents information about  $e$ ; otherwise,  $s$  has the value *unselected*. Variable  $s$  therefore is present and has the value *selected* for each record in  $D_S$ . Note that  $s$  is a variable that is implicit in any observational database; we are making it explicit in order to be able to reason about its influence on the discovery of  $C$  from  $D_S$ .

Wermuth, et al. [23] contains a general discussion of the types of independence and dependence relationships captured by Bayesian networks such as those in Tables 1, 2, and 3. In the current paper, we investigate the extent to which these relationships permit us to distinguish among causal networks from observational data in the presence of selection bias.

The concept of an *unobserved common response variable* is described in [23], and it is suggested that we can represent selection bias using such variables. Thus, according to this line of reasoning, we should classify  $s$  as an unobserved common response variable. In fact, however, in  $D_S$  the variable  $s$  is observed, because for each entity recorded in  $D_S$  we know that the value of  $s$  is *selected*. What generally is unobserved and unknown is the probability distribution of  $s$  (for the total population of interest) conditioned on other variables in  $Z$ .<sup>2</sup> In this paper, we explore the implications for causal discovery from observational data of explicitly modeling the variable  $s$ .

Section 9.3 of Spirtes, et al. [21] contains a general discussion of qualitative and quantitative causal discovery from observational data when a population is sampled according to some particular criteria (e.g., all patients above a certain weight). In the current paper, we assume the sampling criteria is equal to whatever criteria defines the acquisition of the database we are using, which often is implicit and may even be unknown. Thus, we consider criteria that are less specific than one based on particular properties of the sampled entities and that apply without further qualification to all observational databases. We discuss some specific causal scenarios in which such sampling does (and does not) interfere with qualitative and quantitative causal discovery. In [21] random sampling of the population of interest is taken as a general technique for avoiding selection bias when using observational data (at least asymptotically, as the database grows in size); we show cases where this is not so.

### Example

Suppose  $x$  and  $y$  are two rare symptoms (represented as binary variables) that a patient may have, which are caused by two causally independent disease processes.<sup>3</sup> If a patient has either  $x$  or  $y$  alone, then he is very likely to go his family physician. If, however, the patient has both  $x$  and  $y$ , then he is much more likely to go to the emergency room (ER) for fear he is seriously ill. Suppose that we collect observational data  $D_S$  only from patients who visit the ER.  $D_S$  will be subject to a selection bias. In particular,  $s = \textit{selected}$  (i.e., the patient went to the ER, and therefore, was observed) for a each patient represented in  $D_S$ . For the patient records in  $D_S$ , the statistical dependency between  $x$  and  $y$  will be high; in particular, if  $x$  is *present* then it is likely that  $y$  will be *present*, and vice versa. For the entire population of patients in the community, however,  $x$  and  $y$  are marginally independent, because they result from causally independent disease processes. If we use  $D_S$  to suggest causal relationships between  $x$  and  $y$ , we may be misled by the statistical dependency between the two variables that results from selection bias.

There are many possible reasons why selection bias might occur. In the domain of medicine these include (1) a patient's decision about where to be seen, based on symptoms, (2) a physician's decision about where to refer a patient for additional medical care, (3) a clinic's acceptance criteria (e.g., some clinics may only see patients with particular medical conditions), and (4) patient death or

---

<sup>2</sup> We conjecture (but will not pursue here) that each real-world use of an "unobserved" common response variable, involves not an unobserved variable, but rather, an implicit variable  $s$  that has a fixed value (*selected*) in the observational database. What we are missing is data with which to estimate probabilities of the form  $P(s = \textit{unselected} \mid M)$ , where  $M$  is an instantiation of a subset of the model variables. It is these probabilities that are in some sense unobserved.

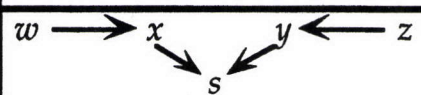
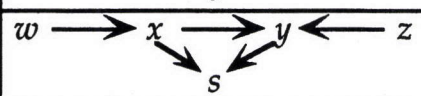
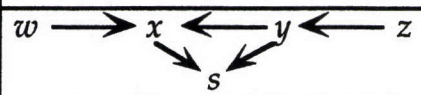
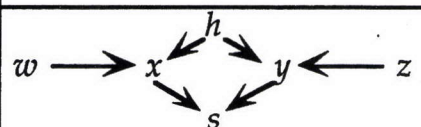
<sup>3</sup> By *causally independent disease processes* we mean that  $x$  and  $y$  share no common ancestors in  $C$ , nor is either the ancestor of the other.



infirmary, which may prevent our ever seeing them. Analogous selection biases can occur in non-medical domains as well, where the entities being observed are not patients.

Table 2 introduces selection variable  $s$  into the causal-network structures taken from Table 1. Note that the key at the bottom of Table 2 shows that the relationships in that table are conditioned on the value of  $s$  being known (as indicated by  $s$  being underlined), which is appropriate, since it is from the observed data  $D_s$  that we infer the independence and dependence relationships in the table, and  $s$  is known to have the value *selected* in  $D_s$ .

Table 2. The value of variable  $s$  indicates whether an entity is selected for observation. In the key given below,  $s$  is underlined as shorthand for " $s = \textit{selected}$ ".

No.	Structure of a causal network $C$	Distinguishing relationships						
		$R_1$	$R_2$	$R_3$	$R_4$	$R_5$	$R_6$	$R_7$
1			+	+	+	+		
2				+	+			
3			+				+	
4								

Key:  $R_1: I_C(x, y | \underline{s})$ ,  $R_2: I_C(x, z | \{\underline{s}, y\})$ ,  $R_3: I_C(w, y | \{\underline{s}, x\})$ ,  $R_4: I_C(w, z | \{\underline{s}, x\})$ ,  
 $R_5: I_C(w, z | \{\underline{s}, y\})$ ,  $R_6: I_C(x, z | \underline{s})$ ,  $R_7: I_C(w, y | \underline{s})$

The following result shows that we can detect the type of selection bias represented by the networks in Table 2.

**Theorem 2.** For  $i$  from 1 through 4, using observational data  $D_s$  it is possible to distinguish causal-network structure  $i$  in Table 2 from the remaining causal-network structures in Tables 1 and 2.

**Proof** The pattern of distinguishing relationships is unique for each causal-network structure in Table 2, relative the other structures in Tables 1 and 2. Thus, if these relationships can be correctly inferred, the structures can be distinguished.  $\square$

Theorem 2 implies that we can sometimes detect the presence of selection bias, and furthermore, such a bias does not interfere with our learning qualitatively the causal relationships between  $x$  and  $y$ . In the presence of such selection bias, however, the following theorem shows that we do not always know how to use  $D_s$  to quantify the relationship between  $x$  and  $y$ . Thus, for example, although we may be able to infer that  $x$  causally influences  $y$  in some way, we cannot learn from just  $D_s$  the likely values of  $y$  that would result from a given manipulation of  $x$ .

**Theorem 3.** For causal-network structure 2 in Table 2, the data in  $D_s$  is insufficient to develop an estimate of the population probability  $P(y | x)$  that converges to that probability as the number of samples in  $D_s$  increases. An analogous situation holds for structure 3 in Table 2.

**Proof** For causal-network structure 2 in Table 2 we can express the probabilistic relationship between  $x$  and  $y$  as  $P(y | x)$ , which we can write as follows:

$$P(y | x) = P(y | x, s = \textit{selected}) P(s = \textit{selected} | x) + P(y | x, s = \textit{unselected}) P(s = \textit{unselected} | x).$$

Database  $D_s$  only contains information relevant to estimating the first term on the right side of the above equation, namely  $P(y | x, s = \textit{selected})$ . The remaining three terms concern information that is outside of database  $D_s$ . Thus, in general,  $D_s$  will not contain sufficient information to derive an



estimate of  $P(y | x)$  that in the limit converges to  $P(y | x)$ . An analogous result holds for structure 3 in Table 2.  $\square$

Thus far, we have seen some situations in which qualitative causal relationships can be learned in the presence of selection bias, even though the discovery of the corresponding quantitative causal relationships is problematic. In some instances, it is possible for selection bias to interfere with the discovery of both qualitative and quantitative causal relationships, as we now show.

Table 3 illustrates three problematic causal networks (there are others as well), which contain a selection variable  $s$  and one or more hidden variables.

Table 3. Nodes  $h, h_1, h_2$  represent hidden variables. In the key,  $s$  is underlined as shorthand for " $s = \text{selected}$ ".

No.	Structure of a causal network C	Distinguishing relationships						
		R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	R <sub>7</sub>
1				+	+		+	
2			+			+		+
3					+	+	+	+

Key: R<sub>1</sub>:  $I_C(x, y | \underline{s})$ , R<sub>2</sub>:  $I_C(x, z | \{\underline{s}, y\})$ , R<sub>3</sub>:  $I_C(w, y | \{\underline{s}, x\})$ , R<sub>4</sub>:  $I_C(w, z | \{\underline{s}, x\})$ ,  
R<sub>5</sub>:  $I_C(w, z | \{\underline{s}, y\})$ , R<sub>6</sub>:  $I_C(x, z | \underline{s})$ , R<sub>7</sub>:  $I_C(w, y | \underline{s})$

Network 1 in Table 3 might, for example, represent an unmeasured genetic condition  $h$  that both influences the likelihood of appearance of symptom  $y$  in patients and influences whether patients are observed (as indicated by variable  $s$ ). For instance, patients may be unlikely to be observed, because the presence of  $h$  increases the likelihood of early patient death. In the extreme, *in utero* death might occur very early during gestation (e.g., at the one cell stage, just after conception), so that there is no possibility of our ever observing an unselected population of people, because they do not exist. In network 1, variable  $x$  also influences whether patients are observed; for example,  $x$  might be a disease that is not causally related to  $h$ . Unfortunately, network 1 in Table 3 has the same distinguishing relationships as network 2 in Table 1. More generally, we have the following result.

**Theorem 4.** For  $i$  from 1 through 3, by using just d-separation relationships derived from observational data  $D_s$ , it is not possible to distinguish causal-network structure  $i$  in Table 3 from causal-network structure  $i+1$  in Table 1.

**Proof** For network structure  $i$  in Table 3 the pattern of its seven distinguishing relationships is identical to the pattern for network structure  $i+1$  in Table 1. Thus, these seven relationships are clearly not adequate to distinguish between these network-structure pairs. For each of the six structures under consideration in Tables 1 and 3, there are a total of 256 d-separation relationships (conditions) on the four measured variables in each structure. Thus, it is possible that structure  $i$  in Table 3 could be distinguished from structure  $i+1$  in Table 1 based on some subset of these 256 relationships that go beyond the seven relationships listed in Tables 1 and 3. A computer program was written that checked all 256 d-separation relationships for each pair of network structures described in the theorem. The results are that all 256 relationships are identical for each pair.  $\square$

According to Theorem 4, there is no way to use just the d-separation relationships that follow from observational data  $D_s$  to uniquely determine the qualitative causal relationships that hold between



$x$  and  $y$ . This situation is not specific to biological systems; in astronomy, for example, an analogous hypothetical scenario could be developed for the death of a star. More broadly, it is possible that our universe itself has been selected from among many possible universes. The parallel universe theory proposed by Everett in 1957 states that each quantum transition splits a universe into parallel (yet mutually inaccessible) universes [6]. Thus, everything that can happen, does happen, in some universe. While at first mention the theory may appear far-fetched, it does explain certain quantum phenomena in a way that seems at least as plausible as other explanations [12]. We can imagine these splitting universes forming a tree, where each branch point in the tree corresponds to a split. According to one line of reasoning (called the anthropic principle) our universe was *selected* because it has properties (e.g., its laws and universal constants) that permit the formation of a place where conscious life can exist [2]. The parallel universe theory suggests that this selection process is happening continually. If correct, these ideas suggest the possibility that selection bias could occur at the level of the universe. Since we do not have access to parallel universes, it would not be possible to avoid such a bias. The implications of such bias, if indeed it exists, have not (as far as we know) been worked out, but they do seem worth pursuing.

Note that Theorem 4 does not exclude the possibility that by measuring additional variables (beyond those shown in Tables 1 and 3) we could distinguish the pairs of structures described in the theorem. We conjecture, however, that it is not possible to ever use observational data to distinguish structure 3 in Table 3 from structure 4 in Table 1, regardless of how many additional observational variables are recorded. If this conjecture is correct, then d-separation information can not be used to distinguish a common hidden cause (structure 4 in Table 1) from separate hidden causes with selection (structure 3 in Table 3). Moreover, experimental studies involving the manipulation of  $x$  (or alternatively the manipulation of  $y$ ) and the measurement of outcome  $y$  (or alternatively the measurement of outcome  $x$ ) cannot distinguish between these two structures, because for both structures the outcome will be statistically independent of the intervention. (See the next section for more discussion of experimental studies.) These limitations suggest the possibility that the only way to establish the presence of a hidden common cause is to observe it, in which case of course it would no longer be hidden.

## Selection Bias in Experimental Studies

Causal discovery from experimental data also can be subject to selection bias. For example, suppose we perform a randomized, controlled trial (RCT) on a set of subjects  $E_S$ . Let  $x = \text{yes}$  denote the experimental intervention (e.g., a drug) that is applied to the experimental group; the control group gets a placebo as the intervention, which we represent as  $x = \text{no}$ . Let variable  $y$  denote the outcome of interest (e.g., the patient has a common cold), which has a value of *yes* or *no*. We measure the value of  $y$  in both the experimental and the control groups. Note first that there may have been some process that selected the subjects available for participation in the experiment (i.e., group  $E_S$ ) from the population of patients about which we want to discover causal relationships (i.e., group  $E_Q$ , as previously defined). Let us assume, however, that such selection bias did not take place. It is still possible that selection bias may exist. In particular, if outcome  $y$  is not measured in all the subjects in  $E_S$ , and is not measured in a uniform fashion (e.g., at the same time and place following the intervention), then selection bias may occur.

For instance, consider network 1 in Table 2. Suppose that the value of  $x$  and the value of  $y$  both determine whether a patient is likely to return to the clinic to have the value of outcome  $y$  measured. In terms of the previous example, suppose that giving the drug (i.e.,  $x = \text{yes}$ ) often causes headaches as a side effect. In a patient the combination of a drug induced headache and a common cold may make it much less likely that the patient will return to the clinic to have the value of  $y$  assessed, than if only one of the two conditions was present. In this case, selection bias would exist, and the statistical dependency between  $x$  and  $y$  might be erroneously taken to mean that  $x$  causally influences  $y$ . We want to make clear that the implications of such "losses to follow" are well recognized and described in the clinical trial literature [14, Chapter 9]. Our purpose is merely to emphasize in the current context that experimental studies are not immune to selection bias.

In situations like the previous example, observational data may allow us to recognize that the dependency between  $x$  and  $y$  is due to selection bias rather than to causation. In particular, if we assume the conditions associated with Tables 1 and 2 hold, then the distinction is possible because network 1 in Table 2 (which indicates selection bias) has distinguishing relationships that are different from network 2 in Table 1 (which indicates causation). Spirtes, et al. [21, Section 9.1]



contains a discussion of other cases in which observational studies can uncover causal relationships that can not be discovered by experimental studies.

In graphical terms, an RCT eliminates the arcs into the intervention node(s), because (ideally) randomization eliminates the possibility that any hidden process is controlling the intervention variable(s) [21, Section 3.7.2]. This removes the possibility of a hidden common cause between the intervention and the outcome. Randomization does not, however, eliminate arcs out of the intervention node, which is the situation that exists when there is selection bias. Thus, experimental studies are vulnerable to many of the same types of selection biases as are observational studies.

As a practical matter, however, it may be easier to avoid selection bias when using an experimental design. Suppose we have enrolled a random sample of a population of patients (call them  $E$ ) about whom we want to discover causal relationships. To avoid selection bias, an experimental study need only make sure that all enrolled subjects are randomly assigned an intervention and that all subjects have their outcomes measured.<sup>4</sup> On the other hand, with observational data it also may be important to know whether some selection process influenced the creation of the subjects in  $E$  — that is, we may need to know something about the history of those subjects. Such historical influences may be totally inaccessible to us. Table 3 illustrates potential selection processes of this type. In particular, recall the *in utero* example (associated with network 1 in Table 3) wherein it is infeasible to observe the selection process that occurs soon after conception. An observational study would encounter selection bias here. An experimental study could in principle avoid selection bias through randomization and complete follow up on the population of subjects of interest. In this example, however, using an experimental study would be unethical, because the intervention (as represented by variable  $x$ ) would require inducing a disease in some subjects. In other situations, experimental studies may be infeasible due to costs or logistical difficulties. Thus observational and experimental studies each have their strengths and weaknesses. Since both types of studies, however, are potentially subject to selection bias, the next section discusses some ways to reduce the influence of such bias on causal inference.

## Approaches to Handling Selection Bias

The results in this paper indicate that in certain circumstances selection bias can hinder our using data to discover causal relationships with any confidence. In this section, we focus on methods for handling selection bias when using observational data for causal discovery.

One approach to addressing this problem is simply to assume that selection bias is nonexistent or sufficiently weak, such that methods using observational data will usually only suggest true causal relationships. Often, however, we may not know the extent to which selection bias exists. Another approach around this problem is to take steps to make sure that we can assume with confidence that selection bias does not exist. For example, we might make our observations from a random sample of the unselected population of interest. In general, however, this may not be practical, due to the effort and expense required to sample from that population. In the worst case, we may not have access to such an unselected population, as the previous example about *in utero* death illustrates. Another way to address the problem would be to perform careful experimental and observational studies on samples of the same entity population, and then empirically determine the extent to which causal discoveries from the observational data predict causal discoveries from the experiments. The better the predictions, the greater our confidence in successfully applying the observational discovery methods to other data about this population.

One other approach to handling selection bias is to try to model it. In some cases, we might, for example, be able to perform limited random sampling of the population of interest in order to model selection bias. Such samples have in fact been taken [18]. It is likely that these samples would influence our subjective prior probabilities regarding selection bias; we could combine these prior probabilities with a (hopefully large) set of observational data (which is subject to the selection bias being modeled) in order to draw conclusions about causal relationships.

In the next section we describe a Bayesian approach to causal discovery that models selection bias explicitly and is based on extensions to the Bayesian causal discovery methods

---

<sup>4</sup> We note that interventions need to be given and outcomes need to be measured in a uniform, consistent fashion across subjects. If not, then selection bias can occur, even if outcome measurements are made for all the subjects.



introduced in [5]. While this approach is not a panacea for all the problems due to selection bias, it does provide us with flexibility in trying to cope with these problems.

## A Bayesian Method for Handling Selection Bias

In this section we sketch a Bayesian method for computing the probability of a causal-network structure that contains a selection variable  $s$ . Key advantages of the Bayesian approach are that (1) it allows us to incorporate prior knowledge about the selection process (e.g., expert opinions), and (2) given prior beliefs and data, we can derive a posterior probability of causal-network structures, in contrast to making a categorical statement about which structure is most likely. The latter point is particularly important when we do not have a large amount of data.

We will focus on the probability of causal-network structure 1 in Table 2. By extension, it is possible to derive methods to compute the probability of the other network structures in Tables 2 and 3, but we will not do so here.<sup>5</sup>

Suppose for the moment that we have a method for calculating  $P(B_{str_i'} | D)$  for some causal belief-network structure  $B_{str_i'}$  and database  $D$ . Let  $Q$  be the set of all those belief-network structures that have a non-zero prior probability. We can derive the posterior probability of  $B_{str_i'}$  given  $D$  as  $P(B_{str_i'} | D) = P(B_{str_i'} D) / \sum_{B_{str} \in Q} P(B_{str} D)$ . Sometimes all we want to know is the ratio of the posterior probabilities of two belief-network structures. To calculate such a ratio for belief-network structures  $B_{str_i'}$  and  $B_{str_j'}$ , we can use the equivalence that  $P(B_{str_i'} | D) / P(B_{str_j'} | D) = P(B_{str_i'} D) / P(B_{str_j'} D)$ .

Our focus in this section will be on computing the term  $P(B_{str_i'} D)$ , which we can derive using the equivalence that  $P(B_{str_i'} D) = P(D | B_{str_i'})P(B_{str_i'})$ . The term  $P(B_{str_i'})$  represents a user's prior probability that a process with belief-network structure  $B_{str_i'}$  generated data  $D$ . We will assume that the user has explicitly specified  $P(B_{str_i'})$ , and it is available. The likelihood term  $P(D | B_{str_i'})$  remains for us to determine. The following theorem provides a method for computing  $P(D | B_{str_i'})$ . A proof of the theorem follows from work reported in Heckerman, et al. [10].

**Theorem 5.** Let  $Z$  be a set of  $n$  variables. Let  $D$  be a database of  $m$  cases, where each case (i.e., record) contains a value assignment for each variable in  $Z$ . Let  $B_{str}$  denote a belief-network structure containing just the variables in  $Z$ . Each variable  $x_i$  in  $B_{str}$  has a set of parents, which we represent with a list of variables  $\pi_i$ . We use  $\text{freq}(x_i | \pi_i)$  to denote the long run frequency distribution of  $x_i$  conditioned on  $\pi_i$ . Let our belief about that long run frequency be represented by a probability distribution over all functions of the form  $\text{freq}(x_i | \pi_i)$ . Let  $\text{OB}(x_i)$  represent our beliefs about  $\text{freq}(x_j | \pi_j)$  for all  $x_j$  such that  $x_j \neq x_i$ . Suppose the following assumptions hold:

1. The variables in  $Z$  are discrete.
2. Cases occur independently, given a belief-network model.
3. There are no cases that have variables with missing values.
4. For each  $x_i$  in  $Z$ , our belief about  $\text{freq}(x_i | \pi_i)$  is not influenced by our other beliefs as given by  $\text{OB}(x_i)$ .
5. For each  $x_i$  in  $Z$ , we represent our belief about  $\text{freq}(x_i | \pi_i)$  using a Dirichlet distribution. In particular, we have a prior belief represented as a Dirichlet distribution (which holds before observing any data) and a posterior belief represented as a Dirichlet distribution (which results from updating our prior using the observed data).

Next, let  $D_j$  designate the first  $j$  cases in  $D$  and let  $C_{j+1}$  denote case  $j+1$  in  $D$ . Consider a belief network  $B_j$  that has structure  $B_{str}$ . The probabilities that parameterize belief network  $B_j$  are derived by taking for each  $x_i$  the expectation of  $\text{freq}(x_i | \pi_i)$  relative to both (1) our prior beliefs about that frequency and (2) the data in  $D_j$ . (See [10] for details regarding how to compute this expectation.)

---

<sup>5</sup> Some of the belief-network structures in Tables 2 and 3 contain hidden variables. To compute the probability of these structures requires the combination of the methods described in this section with methods for handling hidden variables, such as those described [4]. This synthesis is beyond the scope of the current paper, as is the related problem of handling missing data.



Let  $P_{B_j}(C_{j+1})$  represent the probability of the variable instantiations in case  $C_{j+1}$  as computed by belief network  $B_j$ . (See [11] for an overview of algorithms for computing such inferences.)

From the five numbered assumptions above, it follows that

$$P(D \mid B_{str}) = \sum_{j=1, m-1} P_{B_j}(C_{j+1}) \quad (1)$$

□

Let  $B_{str}$  denote network structure 1 in Table 2. Our goal is to derive  $P(D \mid B_{str})$ . Note that  $B_{str}$  is a causal model for the *total population* of entities of interest, and not just those selected for measurement. The problem with using Equation 1 to compute  $P(D \mid B_{str})$  is that assumption 5 in Theorem 5 is violated, because it requires that  $\text{freq}(s \mid \pi_s)$  be revised by calculating a posterior Dirichlet distribution that results from updating a prior distribution with the *measured* data. But,  $\text{freq}(s \mid \pi_s)$  is intended to represent the frequency in the *total* population, which in general contains some entities that are measured and others that are not.

We can address this problem by first dividing the belief network into two submodels. Submodel 1 contains nodes  $w, x, y$ , and  $z$ , the arcs among them in  $B_{str}$ , and the associated parameterizations. Let  $M1_j$  denote this belief network when it is parameterized (as described in Theorem 5) based on a user's prior distribution and the data in cases 1 to  $j$ . Submodel 2 contains nodes  $x, y$ , and  $s$ . Submodel 2 estimates  $P(x, y \mid s = \textit{selected})$ . Let  $M2_j$  denote such a submodel, which is constructed from the user's prior distribution<sup>6</sup> and the data in cases 1 to  $j$ . We do not restrict the method used to derive this estimate, and thus, there are many possibilities. A key point is that we can use measured data to appropriately update our estimate of  $P(x, y \mid s = \textit{selected})$ , whereas measured data gives us no information with which to update our estimate of  $P(s = \textit{selected} \mid x, y)$  beyond our prior for that probability.

Notice that  $x$  and  $y$  d-separate  $w$  and  $z$  from  $s$ , and thus,  $w$  and  $z$  are conditionally independent of  $s$  given  $x$  and  $y$ . We therefore have

$$\begin{aligned} P_{M_j}(C'_{j+1} \mid s_{j+1} = \textit{selected}) &= P_{M1_j}(w_{j+1}, z_{j+1} \mid x_{j+1}, y_{j+1}, s_{j+1} = \textit{selected}) P_{M2_j}(x_{j+1}, y_{j+1} \mid s_{j+1} = \textit{selected}) \\ &= P_{M1_j}(w_{j+1}, z_{j+1} \mid x_{j+1}, y_{j+1}) P_{M2_j}(x_{j+1}, y_{j+1} \mid s_{j+1} = \textit{selected}), \end{aligned} \quad (2)$$

where  $C'_{j+1} = \{w_{j+1}, x_{j+1}, y_{j+1}, z_{j+1}\}$ , and  $P_{M_j}(C'_{j+1} \mid s_{j+1} = \textit{selected})$  designates the probability estimate of case  $C'_{j+1}$  under selection, as computed by the hybrid model. We can apply  $M1_j$  and  $M2_j$  to compute the probability estimates on the right side of Equation 2. By substituting Equation 2 into a modified version of Equation 1, we obtain the following final result:

$$P(D \mid B_{str}, s = \textit{selected}) = \sum_{j=1, m-1} P_{M1_j}(w_{j+1}, z_{j+1} \mid x_{j+1}, y_{j+1}) P_{M2_j}(x_{j+1}, y_{j+1} \mid s_{j+1} = \textit{selected}).$$

## Acknowledgments

I thank Constantin Aliferis and Clark Glymour for helpful comments on an earlier draft of this paper. Discussions with Clark Glymour led me to further appreciate and investigate the degree to which experimental data can be subject to selection bias. This work was supported by grant LM05291-02 from the National Library of Medicine and by grant BES-9315428 from the National Science Foundation.

<sup>6</sup> It may be more natural for a user to specify a prior for  $P(s = \textit{selected} \mid x, y)$ , rather than a prior for  $P(x, y \mid s = \textit{selected})$ . If so, we can derive a prior for the latter from the former by using Bayes theorem.



## References

1. Berkson, J., Limitations of the application of fourfold table analysis to hospital data, *Biometrics* 2 (1946) 47–53.
2. Carr, B.J. and Rees, M.J., The anthropic principle and the structure of the physical world, *Nature* 278 (1979) 605.
3. Conn, H.O., Snyder, N. and Atterbury, C.E., The Berkson bias in action, *The Yale Journal of Biology and Medicine* 52 (1979) 141–147.
4. Cooper, G.F., A method for learning belief networks that contain hidden variables, *Journal of Intelligent Information Systems* 4 (1995) 1–18.
5. Cooper, G.F. and Herskovits, E., A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* 9 (1992) 309–347.
6. DeWitt, B.S. and Graham, N., *The Many-Worlds Interpretation of Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1973).
7. Feinstein, A.R., *Clinical Epidemiology: The Architecture of Clinical Research* (W.B. Saunders, Philadelphia, 1985).
8. Geiger, D., Verma, T. and Pearl, J., Identifying independence in Bayesian networks, *Networks* 20 (1990) 507–534.
9. Gerber, L.M., Wolf, A.M., Braham, R.L. and Alderman, M.H., Effects of sample selection on the coincidence of hypertension and diabetes, *Journal of the American Medical Association* 247 (1982) 43–46.
10. Heckerman, D., Geiger, D. and Chickering, D.M., Learning Bayesian networks: The combination of knowledge and statistical data, In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, (1994) 293–301.
11. Henrion, M., An introduction to algorithms for inference in belief nets. In: Henrion M., Shachter R.D., Kanal L.N. and Lemmer J.F. (eds.), *Uncertainty in Artificial Intelligence* 5 (North-Holland, Amsterdam, 1990) 129–138.
12. Herbert, N., *Quantum Reality* (Anchor Press, Garden City, NY, 1985).
13. Mainland, D., The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease, *American Heart Association* 45 (1953) 644–654.
14. Meinert, C.L. and Tonascia, S., *Clinical Trials: Design, Conduct, and Analysis* (Oxford University Press, New York, 1986).
15. Pearl, J., *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, CA, 1988).
16. Pearl, J., Causal diagrams for empirical research, Report R-218-L, Computer Science Department, UCLA, 1994.
17. Pearl, J. and Verma, T.S., A theory of inferred causality, In: *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, Boston, MA (1991) 441–452.
18. Roberts, R.S., Spitzer, W.O., Delmore, T. and Sackett, D.L., An empirical demonstration of Berkson's bias, *Journal of Chronic Disease* 31 (1978) 119–128.
19. Sackett, D.L., Bias in analytic research, *Journal of Chronic Disease* 32 (1979) 51–63.
20. Spirtes, P., Glymour, C. and Scheines, R., Causality from probability. In: McKee G. (ed.), *Evolving Knowledge in Natural and Artificial Intelligence* (Pitman, London, 1990).
21. Spirtes, P., Glymour, C. and Scheines, R., *Causation, Prediction, and Search* (Springer-Verlag, New York, 1993).
22. Verma, T.S. and Pearl, J., Equivalence and synthesis of causal models, In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, (1990) 220–227.
23. Wermuth, N., Cox, D.R. and Pearl, J., Explanations for multivariate structures derived from univariate recursive regressions, Report 94-1, University of Mainz, 1994.