2016

An overview of causal directed acyclic graphs for substance abuse researchers

Michael Lewis
CUNY Hunter College

Alexis Kuerbis
CUNY Hunter College

Ashdin Publishing
Journal of Drug and Alcohol Research
Vol. 5 (2016), Article ID 235992, 8 pages
doi:10.4303/jdar/235992

Research Article

An Overview of Causal Directed Acyclic Graphs for Substance Abuse Researchers

Michael Lewis and Alexis Kuerbis

Silberman School of Social Work, Hunter College, City University of New York, 2180 Third Avenue, New York, NY 10035, USA
Address correspondence to Alexis Kuerbis, ak1465@hunter.cuny.edu

Received 22 June 2016; Revised 27 July 2016; Accepted 28 July 2016

Abstract *Background.* Within substance abuse research, quantitative methodologists tend to view randomized controlled trials (RCTs) as the "gold standard" for estimating causal effects, in part due to experimental manipulation and random assignment. Such methods are not always possible due to ethical and other reasons. Causal directed acyclic graphs (causal DAGs) are mathematical tools for (1) precisely stating researchers' causal assumptions and (2) providing guidance regarding the specification of statistical models for causal inference with nonexperimental data (such as epidemiological data). *Purpose.* This manuscript describes causal DAGs and illustrates their use in regards to a long standing theory within the field of substance use: the gateway hypothesis. *Design.* Data from the 2013 National Survey of Drug Use and Health are utilized to illustrate the application of causal DAGs in model specification. Then using the model specification constructed via causal DAGs, logistic regression models are used to generate odds ratios of the likelihood of trying heroin, given that one has tried alcohol, marijuana, and/or tobacco. *Conclusion.* Granting the assumptions encoded in specific causal DAGs, researchers, even in the absence of RCTs, can identify and estimate causal effects of interest.

Keywords directed acyclic graphs; DAG; randomized controlled trials; gateway hypothesis

1. Introduction to causal directed acyclic graphs

Experts in quantitative methods, including statisticians, econometricians, and professionals from other disciplines, tend to view *randomized controlled trials* (RCTs) as the "gold standard" for estimating causal effects. The key reason for this is that RCTs utilize random assignment. For example, subjects are randomly assigned to at least two different intervention groups. The purpose of random assignment is to better ensure that these groups are *balanced* on variables, which may have a causal relationship with both the treatment and the outcome of interest.

Variables that affect both the treatment and the outcome of interest are referred to as *confounding variables*. Since randomization aims to generate balance across confounding variables, a researcher who observes a difference between intervention groups on the outcome of interest can be reasonably confident that the difference results from the treatment rather than one or more of the confounding variables.

One major challenge for social science researchers, including those working in the field of addiction, is that some research questions cannot be explored by an RCT in a way that is ethical. While researchers may be interested in estimating causal effects, observational or nonexperimental data appears to limit the ability to do so.

The field of *artificial intelligence* (AI) may provide a solution to such a dilemma. AI researchers have focused on programming computers to "think," and a fundamental feature of thinking relates to causal relationships. Over the years, Pearl [1] and Spirtes et al. [2] have worked on the problem of how to represent thinking about causal relationships in computers by developing mathematical tools for modeling such processes. These same tools can be useful also in the health and social sciences [3].

Such mathematical tools are called *causal directed acyclic graphs* (DAGs). DAGs are useful for researchers interested in estimating causal effects with nonexperimental data for two reasons. First, DAGs provide a way of precisely specifying a researcher's causal assumptions, providing a language to clearly state a researcher's assumptions about what is causing what. By providing this clarity, it allows other investigators to critically evaluate those assumptions—a crucial part of the scientific method. Thus, causal DAGs can serve as an additional resource in a scientific approach to substance use/abuse research.

Second, DAGs provide rules for determining which variables are confounders when a researcher is faced with observational data. These rules then can be used to specify statistical models. For example, DAGs help to determine which variables should and should not be included in a regression model, given the assumptions encoded in the DAG are true.

2. Using the gateway drug hypothesis as an illustration

To better illustrate the abstract ideas discussed below, we refer to a specific example related to substance use/abuse research. The application of DAGs in one test of the *gateway drug hypothesis* (GDH) or gateway theory is described below. The GDH contends that use of certain substances acts as a gateway to the use of other substances [4], usually

in a developmental sequence in which "softer" drugs, such as alcohol and cigarettes, lead to marijuana use, which causes later use of "harder" drugs, such as heroin [5]. Based on its original introduction into the substance abuse literature [6], GDH is used to theorize that cannabis use fosters or facilitates the likelihood of later opioid use. More recent research suggests that this occurs as a result of cannabis altering the opioid system in the brain, thus priming it for later opioid use [7,8].

The GDH is relatively controversial and has generated a good deal of debate. Several studies suggest that instead of a gateway, use across substances is more likely the result of a common cause or a common liability for drug use [5,7]. We do not take a position in this debate. We want simply to use a well-known hypothesis for illustrative purposes.

In this paper, we use a DAG encoding a set of assumptions regarding the GDH to provide an overview regarding the basic ideas involved in DAGs. We illustrate the application of these ideas to model specification using a dataset from the 2013 National Survey of Drug Use and Health (for a description of survey methods see [9]), which was retrieved from the Inter-university Consortium for Political and Social Research. It is important to note that the model we constructed here using DAGs is not the only configuration of the conceptual model one might choose to test the GDH. We chose this specific model and set of relationships on the basis of a combination of our understanding of the literature on the GDH, as well as hunches we have about how certain kinds of drug use are related to others. While some readers might balk at the use of hunches as unscientific, the purpose of DAGS is that any causal assumptions can be represented mathematically—whether based on the literature in a particular area, previous research conducted by an investigator, or an investigator's educated guess—while also providing algorithms for how to address the problem of confounding, granting that the assumptions in question hold. A major purpose of this paper is to illustrate how this process works.

3. The concept of causal effect

We begin a discussion of DAGs with the concept of *causal effect*. According to Chen and Pearl [10], X is a cause of Y if engaging in some course of action to change the value of X would result in a change in the probability distribution of Y. That is, suppose marijuana use is X and heroin use is Y. Assume that both of these variables are binary. In this case, a variable *Marijuana* has the values $0 =$ never tried marijuana or $1 =$ tried marijuana in one's lifetime. Similarly, *Heroin* has the values $0 =$ never tried heroin or $1 =$ tried heroin in one's lifetime.

Suppose there is a group of individuals who have never tried marijuana before; thus they all have a value of 0 for *Marijuana*. Suppose the probability distribution of the *Heroin* values for these individuals is given. Now imagine we do something to change these folks' values of 0 on *Marijuana* to values of 1. If this results in a change in the probability distribution of *Heroin*, then *by definition*, we would have a causal effect.

Pearl has developed a mathematical approach called the *do calculus* to capture the notion of causal effect. Using Pearl's notation, X has a causal effect on Y if $P(Y \mid do(X = x_2))$ is different from $P(Y \mid do(X = x_1))$ [1]. Here x_1 and x_2 are different values of the X variable, P stands for probability, and do captures the idea that the X variable is being set at or forced to take on two different values. This is not the same as simply observing different values of X and comparing the probability distribution of Y at those values. That comparison would amount to comparing $P(Y \mid (X = x_2))$ versus $P(Y \mid (X = x_1))$, which are standard conditional probabilities, and $P(Y \mid do(X = x_2))$ and $P(Y \mid do(X = x_1))$ are not. Instead, they represent how the probability distribution of Y changes *as a result of* some type of *action (or doing)* which deliberately set X at two different values (x_1 and x_2). Applying all this to our ongoing example, if $P(\textit{Heroin} \mid do(\textit{Marijuana} = 1))$ is different from $P(\textit{Heroin} \mid do(\textit{Marijuana} = 0))$, we would have a causal effect of marijuana use on heroin use.

4. Identification versus estimation

The distinction between *identification* and *estimation* is critical for understanding the potential role of DAGs in substance abuse research. Substance abuse researchers likely understand the concept of correlation and its relationship to causality. Two variables, called X and Y, are correlated if X causes Y, Y causes X, or they share a common cause [11]. Elwert states: "identification... determines whether and under what conditions, it is possible to strip an observed association of all its spurious components" (see [12, p. 147]). Thus identification is related to the isolation of "causal correlation" from "noncausal correlation".

Estimation relates to statistical methods, such as OLS regression, to obtain the magnitudes of causal effects. Prior to estimating causal effects, one must identify the causal relationship—in other words determine if causal association can be isolated from noncausal association. Part of the usefulness of causal DAGs is that they provide guidelines for identification (i.e., to isolate causal from noncausal correlation), given that *assumptions made about the causal relationships encoded in a specific DAG are true*. Once such isolation is performed, DAGs help to specify statistical models that estimate the magnitudes of the causal effects.

5. Elements of causal DAGs

Figure 1 was created using *DAGitty*, an open source, online program for drawing causal DAGs [13]. Table 1 defines the variables in Figure 1 and their values. Based on the GDH

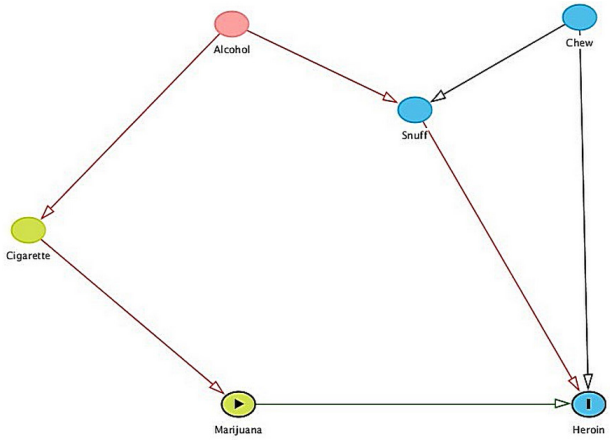

Figure 1: Causal DAG representing causes of having tried *Heroin* in their lifetime.

discussed above, Figure 1 was drawn based on assumptions about the causal relationship between *Marijuana* and *Heroin* and how those two variables are causally related to a set of other variables. The diagram in Figure 1 is an example of a *graph*. A graph is a set of nodes along with arrows connecting those nodes. The set of nodes in Figure 1 is $\{Alcohol, Cigarette, Marijuana, Snuff, Chew, Heroin\}$. The nodes in a causal DAG represent variables, while the arrows represent causal relationships.

5.1. Direct and indirect causal effects depicted by causal DAGs

An arrow "leaving" one variable and "entering" another one represents the assumption that the variable the arrow is leaving causes the variable the arrow is entering. For example, in Figure 1 an arrow leaves *Cigarette* and enters *Marijuana*, and thus the graph encodes the assumption that *Cigarette* causes *Marijuana*. The effect of *Cigarette* on *Marijuana* is also an example of a *direct causal effect*. A direct causal effect is one where the cause variable and the effect variable are separated by just an arrow. That is, there are no other variables between the cause and effect variables.

In Figure 1, not only is *Cigarette* a cause of *Marijuana* but *Marijuana* is also a cause of *Heroin*. Thus, it is the case that *Cigarette* is a cause of *Heroin* by way of its impact on *Marijuana*. This is called an *indirect causal effect*. In general, an indirect causal effect between two variables exists when at least one other variable "stands between" or mediates the cause and effect variables in question. The arrows connecting the cause variable, effect variable, and the mediator(s) must all be, starting from the cause variable, pointing "tail to head". Thus, in this figure, *Cigarette* does not indirectly cause *Snuff*, by way of its effect on *Alcohol*, because the arrow between *Cigarette* and *Alcohol* is pointing head to tail while the one between *Alcohol* and *Snuff* is pointing tail to head.

Table 1: Values of variables from Figure 1.

Variable	Value
Marijuana	1 = Person has tried marijuana in their lifetime 0 = Person has never tried marijuana in their lifetime
Cigarette	1 = Person has tried smoking cigarettes in their lifetime 0 = Person has never tried smoking cigarettes in their lifetime
Alcohol	1 = Person has tried alcohol cigarettes in their lifetime 0 = Person has never tried alcohol in their lifetime
Snuff	1 = Person has tried snuff tobacco in their lifetime 0 = Person has never tried snuff tobacco in their lifetime
Chew	1 = Person has tried chew tobacco in their lifetime 0 = Person has never tried chew tobacco in their lifetime
Heroin	1 = Person has tried heroin in their lifetime 0 = Person has never tried heroin in their lifetime

Table 2: Comparison of path models and causal DAGs.

Path models	Causal DAGs
	Useful for identification
Based on linear relationships	Includes nonlinear relationships
Reciprocal relationships are noted by bidirectional arrows and cycles	Cycles are not allowed. Reciprocal relationships are accounted for by the use of time. Variable X at time 1 affecting variable Y at time 2, affecting X at time 3

At this point, readers familiar with path analysis may conclude that causal DAGs are just another name for path models. This is correct to an extent (see Table 2 for a short comparison). Causal DAGs are a generalization of path models in the following sense. Path models encode linear causal effects. Causal DAGs encode causal effects, which may or may not be linear. That is, an arrow starting from X and ending at Y in a path model means that X has a linear causal effect on Y. Such an arrow in a causal DAG would mean that X has a causal effect on Y without there being a commitment to the form of that effect. This is why causal DAGs are sometimes called "nonparametric" causal models [14].

5.2. Paths

Within causal DAGs a *path* is a sequence of variables connected to each other by arrows [15]. A *directed path* between two variables is one where "travel" is always from the tails to the heads of arrows between variables. These are unidirectional relationships. In Figure 1, the path from *Cigarette* to *Marijuana* to *Heroin* is a directed path between *Cigarette* and *Heroin*.

An *undirected path* between two variables is one where travel along arrows takes place but ignores the direction of the arrows along the path (and thus the direction of the relationship or causal order). In Figure 1, the path from *Cigarette* to *Alcohol* to *Snuff* is an example of an undirected path between *Cigarette* and *Snuff*.

5.3. Cycles

Having defined directed and undirected paths, we can now state an important constraint on the drawing of causal DAGs.

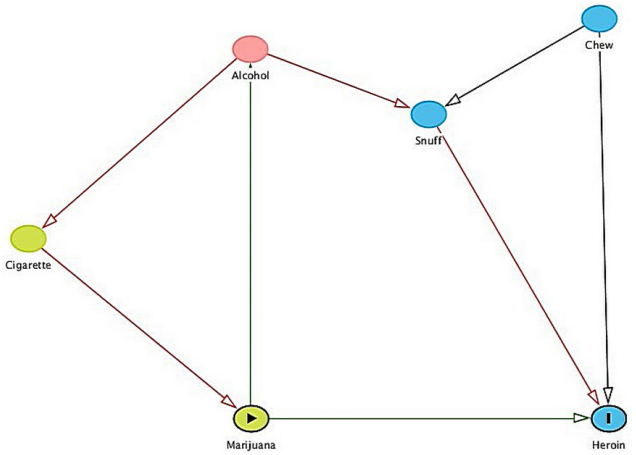

Figure 2: Graph with cycle, representing causes of having tried Heroin in their lifetime.

That constraint is that there can be no cycles. A *cycle* is a directed path that ends with the variable it started with. The directed path in Figure 2 from *Cigarette* to *Marijuana* to *Alcohol* and back to *Cigarette* is an example of a cycle. If a graph contains such a cycle, then that graph, by definition, is no longer a causal DAG. This is because the "A" in DAG stands for "acyclic." Thus, Figure 2 is not a causal DAG.

5.4. Colliders and descendants

Next, the concepts *collider* and *descendant* are important to understanding causal DAGs. When following along a path, a collider is a variable that has two arrows coming into it from two different variables. In Figure 1, consider the path *Marijuana* to *Cigarette* to *Alcohol* to *Snuff* to *Chew* to *Heroin*. A more efficient way of writing out paths is to use notation found in the causal DAGs literature. Using this notation, the path just referred to can be written as *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\leftarrow$ *Chew* $\rightarrow$ *Heroin*. Notice how the directions of the arrows in this notation correspond to the directions of the causal effects in the graph. Also notice that along the path, an arrow enters *Snuff* from *Alcohol* and another one enters *Snuff* from *Chew*. Thus, *Snuff* is a collider. For the path *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin*, *Snuff* is not considered a collider.

A variable Y is a *descendant* of a variable X if there is a directed path from X to Y. In the path *Alcohol* $\rightarrow$ *Cigarette* $\rightarrow$ *Marijuana*, *Marijuana* is a descendant of *Alcohol* because there is a directed path from *Alcohol* to *Marijuana* that includes *Cigarette*.

5.5. Conditioning

Within the context of causal DAGs, *conditioning* occurs when an analyst examines the causal relationship between two variables, *for given values* of at least one other variable, which is the variable on which the relationship is being conditioned. For example, referring to Figure 1, we could examine the causal relationship between *Marijuana* and *Heroin only* for those persons who have tried alcohol in their lifetime (i.e., those who have a value of 1 on the *Alcohol* variable). In this way, we could examine the causal relationship between *Marijuana* and *Heroin*, conditioning on *Alcohol*.

6. Technical relationships between variables in causal DAGs leading to confounding

In this section, concepts related to identification are discussed. For the purposes of this discussion, let us assume that we have a random sample of some population of interest, in which each member has complete data on all the variables in Figure 1. All issues related to lack of a random or probability sample and missing data are eliminated. While important, these matters are not relevant for purposes of this paper. Using these assumptions, we define the following terms: *backdoor path*, *intercepting a path*, *unblocked* (or *open*) *backdoor path*, *blocked* (or *closed*) *backdoor path*, *confounding path*, and *confounders*.

6.1. Backdoor path

Recall that confounding is when a variable affects both the treatment and outcome of interest. From the perspective of causal DAGs, confounding has to do with what is called an unblocked (or open) backdoor path. A *backdoor path* from X to Y is a path which (1) starts from X and ends at Y and (2) has an arrow pointing into X [16]. In other words, when one begins "travel" along a backdoor path, one starts from the head of an arrow and travels towards its tail. In our gateway hypothesis example, let X be *Marijuana* and Y be *Heroin*. The path *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin* in Figure 1 is an example of a backdoor path between *Marijuana* and *Heroin*. This is because (1) the path starts from *Marijuana* and ends at *Heroin*, and (2) there is an arrow pointing into *Marijuana*. To travel along the backdoor path, one would start at the head of the arrow pointing into *Marijuana* and move toward the tail of that arrow.

Researchers familiar with path analysis, but not causal DAGs, might wonder if a backdoor path must have a mediator along it. This is because the path *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin* has a mediator along it if we consider that *Alcohol* causes *Cigarette* (the mediator), which causes *Marijuana*. A backdoor path, however, is not required to have a mediator along it. To see why this is so, consider the following hypothetical examples.

Suppose X causes Y and Z and that Y causes Z. In causal DAG terms, this would be drawn according to Figure 3. There is a backdoor path in the DAG in Figure 3, namely $Y \leftarrow X \rightarrow Z$, because (1) the path starts at Y and ends at Z and (2) there is an arrow pointing into Y; yet, there is no mediator along this path. All we have are direct causes: Y causes Z and X is a common cause of Y and Z.

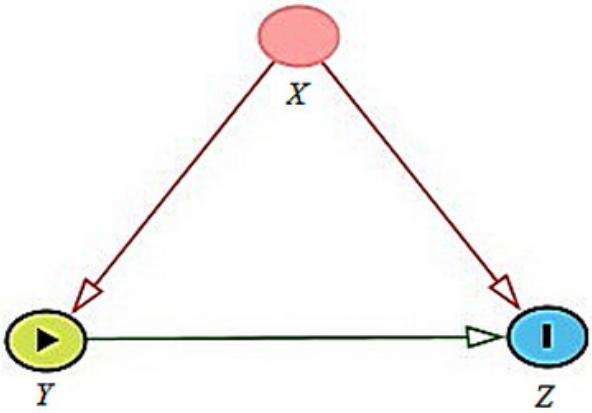

Figure 3: Causal DAG with X common cause of Y and Z.

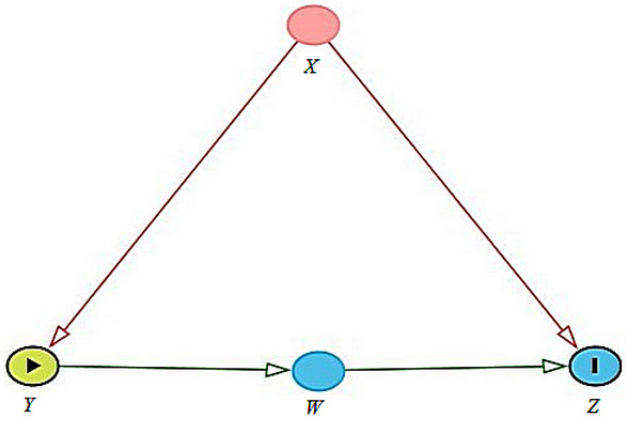

Figure 4: Causal DAG with common cause of Y and Z and W as a mediator.

Now suppose X causes Y which causes W, X causes Z, and W causes Z. Figure 4 demonstrates this in causal DAG terms. There would now be a backdoor path from W to Z because (1) the path starts from W and ends at Z and (2) there is an arrow pointing into W. Since X causes Y, which causes W, W mediates the relationship between X and Y.

6.2. Intercepting a path

A path is *intercepted* by a variable when it is on the path but is not one of the variables at either end of the path [1]. In Figure 4, the path $Y \rightarrow W \rightarrow Z$ is intercepted by W. Additionally, in Figure 1, the path *Alcohol* $\rightarrow$ *Cigarette* $\rightarrow$ *Marijuana* is intercepted by *Cigarette*. A path can be intercepted by more than one variable so long as those variables are on the path but not at either end of it. For example, *Alcohol* $\rightarrow$ *Cigarette* $\rightarrow$ *Marijuana* $\rightarrow$ *Heroin* is intercepted by *Cigarette* and *Marijuana*. A variable that intercepts a path can be, but is not required to be, a mediator.

6.3. Blocked or closed backdoor path

Any backdoor path is considered *blocked* or *closed* if it is intercepted by at least one collider, and *unblocked* or *open* if it is not [16]. In Figure 1, the backdoor path *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\leftarrow$ *Chew* $\rightarrow$ *Heroin* is blocked because it is intercepted by *Snuff*, a collider along that path when it includes *Chew*. The backdoor path *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin* is unblocked because *Cigarette*, *Alcohol*, and *Snuff* are not colliders along that specific path.

6.4. Confounding path and confounders

Within causal DAGs, a *confounding* path is an unblocked or open backdoor path, and the intercepting variables are considered *confounders* [14]. Thus, the (open) backdoor path in Figure 1, *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin* is a confounding path, and members of the set $\{$*Cigarette*, *Alcohol*, *Snuff*$\}$ are confounders along that path.

7. Conditioning on a set of variables to block a confounding path

To identify the causal effect of *Marijuana* on *Heroin*, the model must account for these confounders. To do so, one must block the confounding path. Blocking a confounding path requires conditioning on a set of variables, which is the causal DAGs version of "controlling for" confounders in order to identify a causal effect of interest (see [11, p. 69]).

Suppose **Z** is a set of confounders along a particular confounding path. Conditioning on the variables in **Z** blocks this path if

(1) any variable along the path, which has an arrow leaving it, is a member of **Z** or
(2) the path has at least one collider, which is not a member of **Z**, and no descendant of any collider is a member of **Z** [14].

If the assumptions encoded in a causal DAG are true and there is a set of variables **Z**, which intercepts *all* confounding paths between X and Y, conditioning on **Z** would block all backdoor paths between X and Y. In this way, the causal effect of X on Y is identified by conditioning on **Z** [1].

In Figure 1, there are only two backdoor paths from *Marijuana* to *Heroin*. Notice that the backdoor path (a) *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\leftarrow$ *Chew* $\rightarrow$ *Heroin* is intercepted by *Cigarette*, *Alcohol*, *Snuff*, and *Chew*; while an additional backdoor path (b) *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\rightarrow$ *Heroin* is intercepted by *Cigarette*, *Alcohol*, and *Snuff*.

The backdoor path (a) *Marijuana* $\leftarrow$ *Cigarette* $\leftarrow$ *Alcohol* $\rightarrow$ *Snuff* $\leftarrow$ *Chew* $\rightarrow$ *Heroin* can be blocked without conditioning on anything since *Snuff* is a collider along that path. This would be stated in causal DAG language as conditioning on the empty set $\mathbf{Z} = \{\}$, in which the empty set is one with no members. The empty set contains neither *Snuff* nor a descendent of *Snuff*. This is consistent with the second

requirement above in the definition for blocking a backdoor path. The backdoor path (a) *Marijuana* ← *Cigarette* ← *Alcohol* → *Snuff* ← *Chew* → *Heroin* can also be blocked by conditioning on $\mathbf{Z} = \{Cigarette, Alcohol, Chew\}$ or $\{Cigarette, Alcohol\}$, consistent with the first requirement in the definition for blocking a backdoor path. The backdoor path (b) *Marijuana* ← *Cigarette* ← *Alcohol* → *Snuff* → *Heroin* can be blocked by $\mathbf{Z} = \{Cigarette, Alcohol, Snuff\}$ since the three variables in this set have arrows coming out of them along that path.

Looking closely at *all* the backdoor paths in Figure 1 and which variables block these paths, it becomes clear that conditioning on $\mathbf{Z} = \{Cigarette, Alcohol\}$ blocks *all* the backdoor paths in that DAG. Thus, according to Pearl et al. [14], the causal effect of *Marijuana* on *Heroin* can be identified by conditioning on *Cigarette* and *Alcohol*. The set $\mathbf{Z} = \{Cigarette, Alcohol\}$ is an example of what is called a *sufficient set*. In general, a set of variables is sufficient for identifying the causal effect of X on Y if conditioning on those variables blocks all backdoor paths between X and Y.

The set of variables $\{Cigarette, Alcohol\}$ is a sufficient set, but it is not minimally sufficient. A set of variables is *minimally sufficient* for identifying the causal effect of X on Y if no proper subset of the set is sufficient [1,14] (suppose set $Z_1 = \{a,b,c\}$ and set $Z_2 = \{a,b\}$. Then Z_2 is a proper subset of Z_1 because every member of Z_2 is a member of Z_1 but Z_2 and Z_1 are not equal to or the same as one another). To see this, look again at criterion number 1 for blocking a backdoor path by conditioning on a set of variables along that path: any variable along the path, which has an arrow leaving it. In Figure 1, the variable in set $\mathbf{Z} = \{Alcohol\}$ has an arrow coming out of it along all backdoor paths. So conditioning on it *alone* identifies the causal effect of *Marijuana* on *Heroin*. The same would be true by conditioning on the variable in $\mathbf{Z} = \{Cigarette\}$. Since the sets $\mathbf{Z} = \{Alcohol\}$ and $\mathbf{Z} = \{Cigarette\}$ are both subsets of $\{Cigarette, Alcohol\}$, then $\{Cigarette, Alcohol\}$ is not a minimally sufficient set. Thus, one would not need to condition on both *Alcohol* and *Cigarette* to identify the causal effect of interest. Conditioning on either of them alone would suffice—hence the notion of a minimally sufficient set.

If the causal assumptions encoded in Figure 1 were made, and one further assumed a logit function form, then conditioning on a set of variables could be implemented by including them as "covariates" in a logistic regression model of the causal effect of *Marijuana* on the natural logarithm of the odds (natural log of odds) of *Heroin* (having tried heroin in their lifetime):

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_C\, Cigarette + b_A\, Alcohol.$$

The coefficient b_M would be an estimate of the total causal effect of *Marijuana* on the natural log of the odds of *Heroin*, controlling for the effects of *Cigarette* and *Alcohol* on this outcome. Given the proposed use of logistic regression, we could estimate the effect of *Marijuana* on the probability of having tried heroin in their lifetime, controlling for the effects of *Cigarette* and *Alcohol* on that probability.

When considering minimal sufficiency as described above, the following models could also be run to estimate the causal effect of *Marijuana* on *Heroin*:

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_C\, Cigarette,$$

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_A\, Alcohol.$$

8. Finding confounding paths and sufficient sets

The utility of causal DAGs to substance use/abuse researchers is that once causal assumptions are encoded in a DAG, the DAG provides guidance regarding how to set up regression models to estimate causal effects. Two algorithms relate to using causal DAGs as guides to model specification—one to determine if confounding is present and the other to finding a sufficient (or minimally sufficient) set of variables for inclusion in a model to control for confounding.

For the first algorithm, if a backdoor path between X (cause of interest) and Y (effect of interest) contains the variable Z, which is a common cause of both X and Y, that path is a "candidate" for a confounding path. One can determine if confounding is present by the following steps (see [16] and [11, p. 71]):

(1) delete all arrows coming out of X;
(2) check whether the remaining graph contains variables which cause both X and Y, directly or indirectly;
(3) where there are common causes of X and Y, backdoor paths going through those common causes are confounding paths (unless such a backdoor path goes through a collider or descendant of one). If there are no such common causes of X and Y, then confounding is absent.

Applying this algorithm to the DAG of Figure 1, we would delete the arrow going from *Marijuana* to *Heroin* (step 1), Figure 5. In this scenario, *Alcohol* is the only common cause of *Marijuana* and *Heroin* (step 2), and all backdoor paths going through *Alcohol* would be candidates for confounding paths (step 3). The *Marijuana* ← *Cigarette* ← *Alcohol* → *Snuff* → *Heroin* backdoor path is unblocked so it is a confounding path. The *Marijuana* ← *Cigarette* ← *Alcohol* → *Snuff* ← *Chew* → *Heroin* backdoor path is blocked by *Snuff* since *Snuff* is

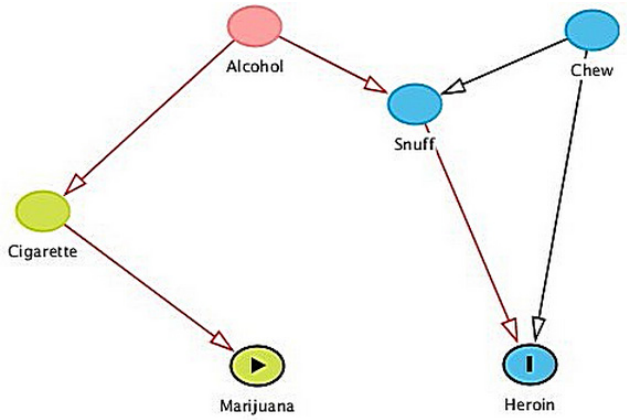

Figure 5: Same as Figure 1 but with arrow between *Marijuana* and *Heroin* deleted.

a collider. To address confounding, all backdoor paths between *Marijuana* and *Heroin* that are not already blocked by *Snuff* must be blocked by conditioning on a sufficient set.

For the second algorithm, in which a sufficient set of variables must be identified for conditioning, the following procedures could be used:

(1) for each backdoor path, put variables that intercept that path into a set;
(2) any set which contains a collider or descendant of a collider blocks the path;
(3) any variable in any set, which is not a collider or descendant of one, can be conditioned on to block the path;
(4) the sufficient set of variables is made up of those that block all backdoor paths between the causal and outcome variables of interest. A minimally sufficient set can be obtained by deleting variables from a sufficient set one at a time until no other variables can be dropped without unblocking the paths between the causal and outcome variables of interest (see [11, p. 72]).

For the DAG in Figure 5, following this second algorithm, $\mathbf{Z} = \{Cigarette, Alcohol\}$ emerges as a sufficient set, and $\mathbf{Z} = \{Cigarette\}$ and $\mathbf{Z} = \{Alcohol\}$ emerge as minimally sufficient sets. It is important to note that while possible, these mathematical algorithms can be tedious to implement by hand, particularly when there may be dozens of variables or paths. *DAGitty* can help to implement the two above described procedures efficiently and effectively.

9. An illustration using data

Based on the algorithms above and the structure of the outcome variable (*Heroin*), we could specify any of the following three logistic regression models:

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_C\, Cigarette + b_A\, Alcohol; \quad (1)$$

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_C\, Cigarette; \quad (2)$$

$$\ln\left(\frac{P(Heroin=1)}{1-P(Heroin=1)}\right) = a + b_M\, Marijuana + b_A\, Alcohol. \quad (3)$$

That is, given the causal relationships encoded in the DAG in Figure 1, as well as the algorithms discussed above, each of these models could be used to estimate the causal effect of *Marijuana* on *Heroin*. In the "real world", the decision regarding which equation to use could depend on data availability. Fortunately, within the 2013 National Survey of Drug Use and Health (retrieved from the website of the Inter-university Consortium for Political and Social Research) [9], all the variables referred to in the DAG of Figure 1 were available for approximately 55,160 participants in this wave of data. Thus, we ran all three models which resulted in three different estimates of the causal effect of *Marijuana* on *Heroin*; however, all of these estimates would be considered equivalent, apart from sampling variation, which stems from the fact that a finite sample was taken from a population [15].

Our findings for models (1)–(3) above were 20.3 (13.2, 31.0), 22.8 (14.9, 34.9), and 36.1 (23.5, 55.4). The first numbers listed are adjusted odds ratios while the numbers in parentheses are confidence intervals for those ratios. For example (model (2)), controlling for the effect of having tried cigarettes, having tried marijuana is estimated to cause the odds of having tried heroin to be about 23 times that of persons who have not tried marijuana. None of the three confidence intervals contains 1; thus, assuming the causal assumptions encoded in the DAG of Figure 1 are true, we appear to have support for the gateway drug hypothesis. These findings are consistent with existing literature examining the veracity of the GDH in American samples [7].

10. Conclusion

Substance use/abuse researchers often want to make causal inferences with observational data. Causal DAGs are useful tools in this effort, particularly with epidemiological data, in helping to provide precise language in visualizing causal assumptions. A causal DAG is most useful if a researcher already has strong assumptions about what is causing what, such as in those areas with developed theories. We looked at such a theory in this paper—the gateway drug hypothesis, and in an illustration of the use of causal DAGs for model specification, we found strong support for it (causal DAGs have been used for causal discovery as well as to guide model specification. By "causal discovery" we mean starting with data and employing algorithms to "search" those data for the causal DAGs which generated

them. This use of DAGs is more controversial than the one we have focused on in this paper. See [2] for further discussion).

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] J. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, New York, 2000.

[2] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, MIT press, Cambridge, MA, 2nd ed., 2000.

[3] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern Epidemiology*, Wolters Kluwer Health, Lippincott Williams & Wilkins, New York, 3rd ed., 2008.

[4] K. Bell and H. Keane, *All gates lead to smoking: the 'gateway theory', e-cigarettes and the remaking of nicotine*, Soc Sci Med, 119 (2014), 45–52.

[5] D. Kandel and E. Kandel, *The Gateway Hypothesis of substance abuse: developmental, biological and societal perspectives*, Acta Paediatr, 104 (2015), 130–137.

[6] D. Kandel and R. Faust, *Sequence and stages in patterns of adolescent drug use*, Arch Gen Psychiatry, 32 (1975), 923–932.

[7] L. Degenhardt, L. Dierker, W. T. Chiu, M. E. Medina-Mora, Y. Neumark, N. Sampson, et al., *Evaluating the drug use "gateway" theory using cross-national data: consistency and associations of the order of initiation of drug use among participants in the WHO World Mental Health Surveys*, Drug Alcohol Depend, 108 (2010), 84–97.

[8] M. Ellgren, S. M. Spano, and Y. L. Hurd, *Adolescent cannabis exposure alters opiate intake and opioid limbic neuronal populations in adult rats*, Neuropsychopharmacology, 32 (2007), 607–615.

[9] Substance Abuse and Mental Health Services Administration, *Results from the 2013 National Survey on Drug Use and Health: Summary of National Findings*, NSDUH Series H-48, HHS Publication No. (SMA) 14-4863, Substance Abuse and Mental Health Services Administration, Rockville, MD, 2014.

[10] B. Chen and J. Pearl, *Regression and causation: a critical examination of six econometrics textbooks*, Real-World Economics Review, 65 (2013), 2–20.

[11] M. A. Lewis, *An overview of causal directed acyclic graphs for social work researchers*, J Appl Quant Methods, 10 (2015), 60–76.

[12] F. Elwert, *Graphical causal models*, in Handbook of Causal Analysis for Social Research, S. L. Morgan, ed., Springer-Verlag, New York, 2013, 245–273.

[13] J. Textor, J. Hardt, and S. Knüppel, *DAGitty: a graphical tool for analyzing causal diagrams*, Epidemiology, 22 (2011), 745.

[14] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer*, John Wiley & Sons, United Kingdom, 2016.

[15] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge University Press, New York, 2nd ed., 2014.

[16] S. Greenland, J. Pearl, and J. M. Robins, *Causal diagrams for epidemiologic research*, Epidemiology, 10 (1999), 37–48.