Paper SAS2998-2019

Causal Graph Analysis with the CAUSALGRAPH Procedure

Clay Thompson, SAS Institute Inc., Cary, NC

ABSTRACT

Valid causal inferences are of paramount importance both in medical and social research and in public policy evaluation. Unbiased estimation of causal effects in a nonrandomized or imperfectly randomized experiment (such as an observational study) requires considerable care to adjust for confounding covariates. A graphical causal model is a powerful and convenient tool that you can use to remove such confounding influences and obtain valid causal inferences. This paper introduces the CAUSALGRAPH procedure, new in SAS/STAT® 15.1, for analyzing graphical causal models. The procedure takes as its input one or more causal models, represented by directed acyclic graphs, and finds a strategy that you can use to estimate a specific causal effect. The paper provides details about using directed acyclic graphs to represent and analyze a causal model. Specific topics include sources of association and bias, the statistical implications of a causal model, and identification and estimation strategies such as adjustment and instrumental variables. Examples illustrate how to apply the procedure in various data analysis situations.

INTRODUCTION

In an experimental setting, the effect of an intervention (for example, a drug treatment or a public policy) can be investigated by randomly assigning experimental units (for example, individuals or households) to either a treatment group or a control (untreated or unexposed) group. Then the magnitude of the causal effect can be estimated by directly comparing the measured outcomes in the two groups. This estimate has a valid causal interpretation because the randomization step enables you to safely assume that no confounding variables are associated with both the treatment assignment and the outcome.

However, in many situations the randomization of units into treatment and control groups is either impractical or unethical. In these cases, you must rely on observational data in order to estimate a causal effect. Causal analysis of observational data, in contrast to data from randomized experiments, requires great care to identify possible confounding variables that are associated with both treatment and outcome, and to adjust for the bias that is created by these variables.

A causal graph is a powerful, easy-to-use tool that you can use to analyze the relationships among treatment variables, outcome variables, and other covariates. A causal graph is created when a causal model is encoded in the form of a directed acyclic graph (Pearl 2009a, b) that depicts the assumed causal relationships in a data generating process. The CAUSALGRAPH procedure, new in SAS/STAT 15.1, examines the structure of causal graphs and suggests statistical strategies or steps that enable you to estimate causal effects that have valid causal interpretations. This paper reviews the role of causal graphs in the analysis of observational data and includes examples of how you can use the CAUSALGRAPH procedure to help conduct a valid causal analysis.

The next section uses a small example to demonstrate the care required when performing a causal analysis that uses observational data. That section is followed by a larger, more practical example of the CAUSALGRAPH procedure. Next, the section "Main Features of the CAUSALGRAPH Procedure" summarizes the capabilities of the procedure. This is followed by the section "Theory of Causal Graph Analysis," which establishes the definitions and theoretical foundations on which the CAUSALGRAPH procedure is based. This theoretical section contains technical details that you can skip in a first reading. Finally, the paper concludes with several examples of the use of the CAUSALGRAPH procedure. Complete code for all examples in this paper is available online.[1]

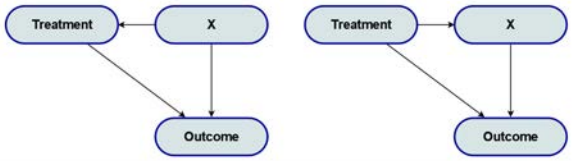
MOTIVATION FOR GRAPHICAL MODELS IN CAUSAL ANALYSIS

In order to understand some of the practical difficulties that arise when you use observational data to estimate a causal effect, consider the two simple data generating processes that are represented graphically in Figure 1. Both processes contain three variables: **Treatment**, **Outcome**, and a covariate **X**. Also in both cases, **Treatment** plays

[1] http://support.sas.com/rnd/app/stat/examples/sgf19_causalgraph.html

a direct causal role in determining the value of **Outcome**, as is indicated by the directed edge that links **Treatment** to **Outcome**. The meaning and interpretation of causal graphs is discussed in more detail in the section "Theory of Causal Graph Analysis."

Figure 1 A Confounding Covariate (Left) and a Mediating Covariate (Right)



The two processes differ in terms of the role of the covariate. In the left panel of Figure 1, **X** is a common cause of both **Treatment** and **Outcome**. One example of such a data generating process would be if **Treatment** is a binary variable that indicates whether a person is taking a particular drug and **Outcome** is a binary variable that indicates whether a person has experienced a particular side effect. If men are more likely than women to be prescribed the drug and also more likely to experience the side effect, then gender is a common cause of **Treatment** and **Outcome** and plays the role of **X** in the left panel of Figure 1.

In the right panel of Figure 1, **X** is caused by **Treatment**, and **X** in turn causes **Outcome**. In other words, **X** mediates a portion of the causal effect of **Treatment** on **Outcome**. As an example of such a data generating process, consider a variation of the example in the preceding paragraph in which the drug can cause a change in blood pressure and this change then increases the probability of the side effect (say, fainting). In this case, blood pressure mediates part of the effect of **Treatment** on **Outcome** and plays the role of **X** in the right panel of Figure 1.

Now, assume you have data for all three variables from an observational study and you want to quantify the magnitude of the causal effect of **Treatment** on **Outcome**. This leads to a seemingly simple question: should you adjust for the covariate **X** in order to estimate the effect? This decision has important consequences for the accuracy of the estimate because adjustment can change not only the magnitude of the estimated effect but even the sign of the causal effect (Pearl 2014).

Intuitively, in the left panel of Figure 1 the variable **X** creates association between **Treatment** and **Outcome** because it plays a role in causing both. However, the association between **Treatment** and **Outcome** that is the result of their common cause is not a part of the causal association that you are interested in computing. In other words, some of the association between **Treatment** and **Outcome** is explained by the biasing path that runs from **Treatment** through **X** to **Outcome**. In order to compute the causal effect, this biasing path must be blocked by adjusting for **X**.

Meanwhile, in the right panel of Figure 1, the variable **X** is a part of the total causal effect of **Treatment** on **Outcome**. In other words, some of the association between **Treatment** and **Outcome** is explained by the causal path that runs from **Treatment** through **X** to **Outcome**. This means that you should not adjust for the variable **X** when computing the causal effect, or else you would block a causal path.

Notice that the adjustment decision is based entirely on a representation of a data generating process in the form of a graphical causal model. It is important to realize that the causal model associated with a set of variables represents a set of assumptions about the data generating process. Although the assumptions that form the causal model do have implications that can be tested in the data (see the section "Theory of Causal Graph Analysis"), these assumptions generally cannot be determined from the data. Rather, you must justify the causal model on the basis of expert knowledge, prior analysis, and so on. You can then use this causal model to determine whether it is possible to estimate a causal effect from observational data and, if so, how to estimate that effect.

Rather than relying on intuition as in the preceding discussion, you can use the CAUSALGRAPH procedure, new in SAS/STAT 15.1, to help you decide which covariates should be included in the adjustment. This procedure is particularly useful in situations where causal models are more complicated and unmeasured variables are involved. You can use the following code to examine the effect of adjusting for the covariate **X** in the two causal models shown in Figure 1:

```
proc causalgraph;
   model "Confounding Covariate" Treatment ==> Outcome, X ==> Treatment Outcome;
   model "Mediating Covariate" Treatment ==> Outcome, Treatment ==> X ==> Outcome;
   identify Treatment ==> Outcome;
```

```
   testid "No Adjustment";
   testid "Covariate Adjustment" X;
run;
```

Each of the causal models shown in Figure 1 is defined in a separate MODEL statement. The causal effect of interest is specified in the IDENTIFY statement. Each of the TESTID statements specifies an adjustment set (a null set and the singleton set **X**) and requests an analysis of whether that set can be used to obtain a valid estimate of the causal effect for each of the causal models.

The analysis results are shown in Figure 2 and Figure 3. As expected, the null set (Figure 2) is a valid adjustment for the mediating covariate causal model, but it is not a valid adjustment for the confounding covariate causal model. Conversely, adjustment for the covariate (Figure 3) is valid for the confounding model but not for the mediating model.

Figure 2 Test Results for the Null Adjustment Set Applied to the Causal Models in Figure 1

Covariate Adjustment Test: No Adjustment
Causal Effect of Treatment on Outcome

Model	Size	Valid	Minimal	Covariates X
Confounding Covariate	0	No	No	
Mediating Covariate	0	Yes	Yes	

Figure 3 Test Results for the Covariate Adjustment Set Applied to the Causal Models in Figure 1

Covariate Adjustment Test: Covariate Adjustment
Causal Effect of Treatment on Outcome

Model	Size	Valid	Minimal	Covariates X
Confounding Covariate	1	Yes	Yes	*
Mediating Covariate	1	No	No	*

The CAUSALGRAPH procedure analyzes causal models and suggests identification strategies (such as which variable to include in an adjustment set) that you can use to obtain unbiased estimates of causal effects. This brings to mind several questions. First, how can you create causal graphs that represent a data generating process? Second, how does the CAUSALGRAPH procedure construct identification strategies? Third, can the CAUSALGRAPH procedure be used for more practical examples, such as those involving many covariates, multiple paths, and unmeasured variables? The first two questions are addressed in the section "Theory of Causal Graph Analysis." The (affirmative) answer to the third question is demonstrated through several examples.

EXAMPLE 1: CAUSAL EFFECT IDENTIFICATION WITH PROC CAUSALGRAPH

The example in the previous section uses two very simple data generating processes to demonstrate the care that is required when you attempt to estimate causal effects from observational data. As shown in that example, you use covariate adjustment in order to block noncausal association between a treatment variable and an outcome variable. On the other hand, if the covariate mediates a part of the causal association, then you should avoid adjusting for that variable so that you do not block part of the causal association.

Although this justification is intuitive, most practical examples are too large to be considered in this way. Rather, you need a software tool that can analyze a causal graph and suggest statistical strategies that can be used in order to compute an unbiased estimate of a causal effect. The CAUSALGRAPH procedure can perform such an analysis, as demonstrated in the following example.

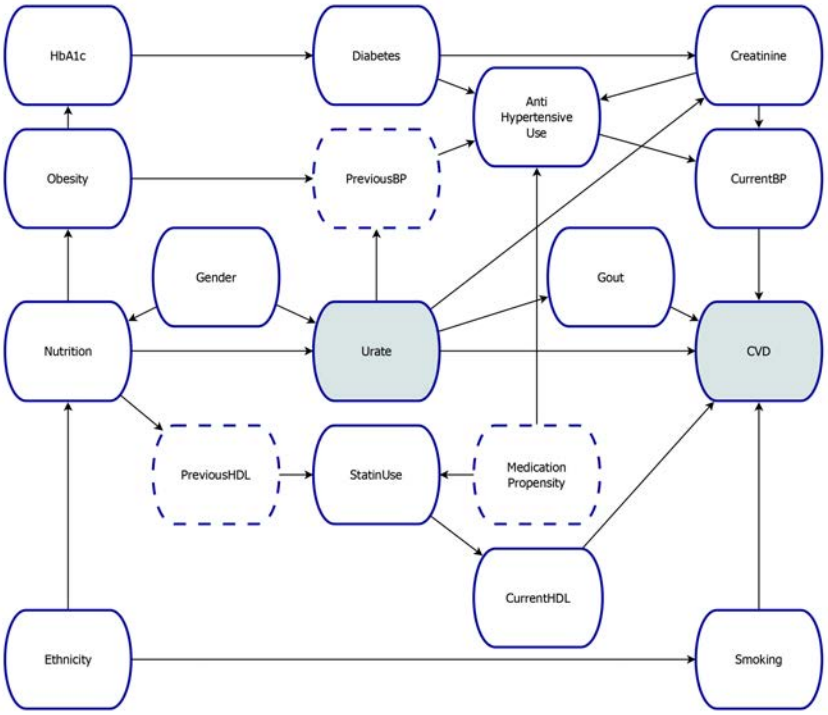
The causal model shown in Figure 4 is adapted from Thornley et al. (2013), where it was used to examine the relationship between an individual's serum urate and risk of cardiovascular disease. The model includes the following variables:

- **Urate**: the treatment variable
- **CVD**: the outcome variable

- **AntiHypertensiveUse**: indicator of antihypertensive drug use
- **Creatinine**: measured serum creatinine level
- **Diabetes**: indicator of diabetes diagnosis
- **Ethnicity**: classification variable for ethnicity
- **Gender**: indicator for biological male
- **Gout**: indicator of gout diagnosis
- **HbA1c**: measured glycated hemoglobin
- **MedicationPropensity**: latent construct that reflects an individual's propensity to take prescribed medication
- **Nutrition**: latent construct that reflects diet or nutrition
- **Obesity**: indicator of body mass index ≥ 30
- **CurrentBP**: measured blood pressure
- **CurrentHDL**: measured HDL cholesterol
- **PreviousBP**: previous (prior to study) blood pressure
- **PreviousHDL**: previous (prior to study) HDL cholesterol
- **Smoking**: indicator of current smoking status
- **StatinUse**: indicator of statin drug use

The variable **MedicationPropensity** corresponds to a latent construct and thus cannot be observed. It is also assumed that the variables **PreviousBP** and **PreviousHDL** are not observed. The unobserved variables are indicated by broken outlines in Figure 4.

Figure 4 Causal Model of the Effect of Serum Urate on Risk of Cardiovascular Disease



In the causal model in Figure 4, some of the association between the variables **Urate** and **CVD** is attributable to several causal paths. For example, it is assumed that there is a direct causal effect of **Urate** on **CVD**. It is also assumed that there are several indirect causal effects, such as the path from **Urate** to **Gout** to **CVD**. Finally, some of the association between the variables **Urate** and **CVD** is assumed to be attributable to noncausal pathways. For example, **Ethnicity**, **Nutrition**, and **Smoking** together create a confounding association between **Urate** and **CVD**.

In a causal analysis, you must determine whether it is possible to isolate and remove the noncausal association between a treatment and an outcome. One possible strategy for doing so is to use covariate adjustment. You can use the CAUSALGRAPH procedure to determine which covariates to include in an adjustment set in order to estimate the total effect of **Urate** on **CVD**. To do so, you use the following code:

```
proc causalgraph compact minimal;
   model "Thor12"
      AntiHypertensiveUse ==> CurrentBP,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      CurrentBP ==> CVD,
      CurrentHDL ==> CVD,
      Diabetes ==> AntiHypertensiveUse Creatinine,
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      HbA1c ==> Diabetes,
      MedicationPropensity ==> AntiHypertensiveUse StatinUse,
      Nutrition ==> PreviousHDL Urate Obesity,
      Obesity ==> PreviousBP HbA1c,
      PreviousBP ==> AntiHypertensiveUse,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL,
      Urate ==> PreviousBP Creatinine CVD Gout;
   identify Urate ==> CVD;
   unmeasured PreviousBP PreviousHDL MedicationPropensity;
run;
```

In the MODEL statement, you specify the causal model to be analyzed. The quoted string in the statement labels the model, and the remainder of the MODEL statement specifies all the variables and edges in the model. These variables and edges encode the hypothesized data generating process that is shown in Figure 4.

In the IDENTIFY statement, you specify the causal effect of interest. In this example, you are interested in the causal effect of **Urate** on **CVD**. In the UNMEASURED statement, you specify variables that are not observed and thus cannot be included in an adjustment set. In this example, three variables are specified as unmeasured.

The MINIMAL option in the PROC CAUSALGRAPH statement requests only those adjustment sets that cannot be made smaller while remaining valid adjustments. The COMPACT option displays the output table in a compact manner.

The results of the CAUSALGRAPH analysis are shown in Figure 5. There are three valid minimal adjustment sets for the model shown in Figure 4. Under the assumption that the causal graph in Figure 4 is an accurate representation of the true data generating process, you can use any one of these three adjustment sets in order to estimate the causal effect of **Urate** on **CVD**. For more information about how to use adjustment sets to estimate a causal effect, see the section "Theory of Causal Graph Analysis." For an example, see "Example 2: Estimating a Causal Effect."

Figure 5 Valid Minimal Adjustment Sets for the Model in Figure 4

Covariate Adjustment Sets for Thor12

Causal Effect of Urate on CVD

			Covariates			
	Size	**Minimal**	**Ethnicity**	**Gender**	**Nutrition**	**Smoking**
1	2	Yes	*		*	
2	2	Yes		*	*	
3	2	Yes			*	*

MAIN FEATURES OF THE CAUSALGRAPH PROCEDURE

This section summarizes the most important features of the CAUSALGRAPH procedure.

Inputs to the CAUSALGRAPH Procedure

The primary input to the CAUSALGRAPH procedure is one or more causal models in the form of directed acyclic graphs (DAGs). In the context of causal models, a DAG is a collection of nodes (variables) and edges (arrows) that define the assumed causal relationships in a data generating process. For more information about the use of DAGs to represent a causal model, see the section "Theory of Causal Graph Analysis."

In the CAUSALGRAPH procedure, you use a MODEL statement to specify a DAG. The MODEL statement supports a pathlike syntax to input causal relationships among variables. For example, to specify the causal path $X \rightarrow Y$, you can use either the **`X ==> Y`** or **`Y <== X`** syntax in the MODEL statement. You can also specify multiple causal relationships as a chain of causal paths: for example, **`X ==> Y ==> Z`**, **`Z <== X ==> Y <== W`**, and so on.

You can specify variables that are not measured by using the UNMEASURED statement. Any variable that is not included in the UNMEASURED statement is assumed to be measured or observed. You can use only measured variables to estimate the causal effect.

The CAUSALGRAPH procedure supports the specification of bidirected edges. A bidirected edge syntax, such as **`X <==> Y`** (for $X \leftrightarrow Y$), is interpreted as unmeasured confounding between the two variables, so that the graph is still a DAG. That is, **`X <==> Y`** is equivalent to **`X <== L ==> Y`** (for $X \leftarrow L \rightarrow Y$), where the node L represents some unmeasured variable.

After you have specified the theoretical data generating model, you can state the causal effect of interest by using the IDENTIFY statement to specify a set of treatment variables and a set of outcome variables. When you specify multiple treatment variables, the CAUSALGRAPH procedure interprets the causal effect as a joint (that is, simultaneous) treatment effect. You cannot use the CAUSALGRAPH procedure to analyze a dynamic treatment regime. When you specify multiple outcome variables, the CAUSALGRAPH procedure examines whether it is possible to use a single identification strategy (for example, a single adjustment set) to estimate the effect of the treatment variables on each outcome variable.

Methods for Causal Effect Identification

A causal effect is said to be identified if it is possible to use the available data to construct an estimate of the effect that has a valid causal interpretation. For more information about causal effect identification, see the section "Theory of Causal Graph Analysis." You can specify the following methods for causal effect identification by using the METHOD= option in the PROC CAUSALGRAPH statement:

- constructive backdoor criterion method (Van der Zander, Liśkiewicz, and Textor 2014) (METHOD=ADJUSTMENT)
- backdoor criterion method (Pearl 2009b) (METHOD=BACKDOOR)
- instrumental variables method (Van der Zander, Textor, and Liśkiewicz 2015) (METHOD=IV)

The first two methods use adjustment-based criteria for identifying a causal effect. The constructive backdoor criterion method is complete for covariate adjustment. This means that if there is any set of covariates that can be used as a valid adjustment set to estimate a causal effect, this method can locate that set. Because of this completeness property, the constructive backdoor criterion method is typically the first method you should use if you want to determine whether a causal effect can be identified. This method is also the default if you do not specify the METHOD= option. The backdoor criterion method is very similar to the constructive backdoor criterion method. It is a popular adjustment method as a result of its intuitive graphical interpretation of adjusting for covariates that explain the assignment into treatment (Elwert 2013). In a causal diagram, these covariates appear in paths that contain an edge that points to a treatment variable. Although the backdoor criterion method is also computationally more efficient than the constructive backdoor criterion method, it is not complete for adjustment. This means that even if the backdoor criterion method fails, it might still be possible to identify a causal effect by using an adjustment set.

The instrumental variables method is an alternative technique for identification that does not rely upon adjustment. This method can be useful when latent or unobserved confounding is believed to exist between a treatment variable and an outcome variable, which means that you cannot use adjustment to correct for the confounding association.

Other Important Features

PROC CAUSALGRAPH has two primary modes of operation:

- Listing mode: The procedure enumerates the criteria under which the causal effect is identified. This is the default mode, or you can specify this mode by using the LIST option.

- Testing mode: The procedure tests whether a user-specified identification criterion is valid. This mode operates when you use the TESTID statement to specify the identification criterion.

You can use both of these modes in a single run of the procedure.

Although a causal model is based on assumptions about the data generating process and cannot be determined from the data alone, a causal model does have implications for the statistical properties in any data set that is consistent with the causal model. For more information about these statistical properties, see the section "Theory of Causal Graph Analysis." You can use the IMAP option in the PROC CAUSALGRAPH statement to request an enumeration of these properties. This can be a useful tool to build confidence that a causal model is consistent with an underlying data generating process. It can also provide a useful method to distinguish between competing causal models on the basis of which is a better representation of the data. For an example of the IMAP option as a tool to distinguish between alternative causal models, see "Example 4: Distinguishing between Alternative Models."

THEORY OF CAUSAL GRAPH ANALYSIS

This section describes the definitions and theoretical foundations on which the CAUSALGRAPH procedure is based. It contains technical details that you can skip in your first reading.

In order to estimate a causal effect by using data from a nonrandomized study, you must supplement the data with a set of causal assumptions. These assumptions collectively form a causal model. Once a causal model has been defined, you can use a set of algorithmic tools to determine strategies for unbiased estimation of the causal effect. These algorithmic tools are implemented in the CAUSALGRAPH procedure.

You should always keep in mind that all inferences, estimates, and conclusions that are drawn from a causal model are contingent on the accuracy of the causal model. For example, the estimation strategies suggested by the CAUSALGRAPH procedure are "unbiased" only under the assumption that the input causal model is an accurate representation of the true data generating process. One of the strengths of graphical models for causal inference is that the causal graph provides a concise tool to represent the assumptions of a causal model. When defining a causal graph, you should carefully consider the causal assumptions that are encoded in the graph and defend these assumptions on the basis of expert knowledge, prior experience, and so on. The next subsection provides information on the manner in which causal assumptions are encoded in a causal graph.

Causal Graphs as Nonparametric Structural Equation Models

The primary input to the CAUSALGRAPH procedure is a causal model in the form of a directed acyclic graph (DAG). Each node in a DAG represents a variable that is assumed to play a role in the data generating process that you want to study. It is not necessary for every node in the DAG to correspond to a variable that has been (or could be) measured. For example, some variables in the DAG might correspond to latent constructs that are assumed to play a causal role in the process being modeled but that cannot be observed directly.

Nodes in the DAG are joined together by edges, which are directed arrows that point from one node to another node. Each edge represents a causal assumption. For example, the edge $X \rightarrow Y$ represents the assumption that the variable X has a possible direct causal role in determining the value of the variable Y. Because each edge in a DAG is given a causal interpretation, each edge is associated with a temporal ordering of a pair of nodes. For this reason, the DAG cannot contain a directed cycle.

The strongest assumptions made in a causal model occur in the form of edges that are *missing* from the DAG. Specifically, if there is no edge that directly connects two variables U and V in a causal graph, then these variables are assumed to have no direct influence on each other. This is the the *strong null hypothesis* for graphical models, and it corresponds to an *exclusion restriction* in econometrics literature (Elwert 2013).

A causal graph is fully nonparametric in the sense that no assumption is made regarding the distribution function of the variables included in the graph and no assumption is made regarding the functional form of the edges that are included in the graph. The only requirement is the assumption that the value of each variable is uniquely determined by the values of its parents and any exogenous (disturbance or error) terms. In this sense, a causal graph can be interpreted as a form of nonparametric structural equation model (NPSEM) (Elwert 2013). In other words, the DAG defines the causal relationships that determine how the value of each variable in the model is determined. Your beliefs about these relationships reflect an existing state of knowledge about both the subject matter being studied and the measurement process used for data collection.

By convention, error random variables that are assumed to play a causal role in determining the value of a single modeled variable are not included in the DAG. However, any variable (including random disturbances) that affects two or more modeled variables must be included in the causal model. You cannot omit any confounding variable from the causal model.

Although a DAG provides a formal semantics for defining a causal model, causal analysis that is based on a DAG is valid only if the DAG (interpreted as a NPSEM) is an accurate representation of the true data generating process. The good news is that a DAG encodes certain statistical implications for the data generating process, and these implications can be tested in the available data. This is discussed in the next subsection.

Association and Bias in Causal Graphs

A DAG encodes structural relationships between the variables in a causal model. These structural relationships between variables in a causal model have implications for the statistical associations between variables in a data set. You can think of these implications as reflecting the flow of information in the data generating process. This information flow is determined by three fundamental structures (Elwert 2013) that are summarized in Table 1.

Table 1 Three Fundamental Structures in a Directed Acyclic Graph

Path	Association between U and W along the Path	Role of V in the Path	Effect of Conditioning on V
$U \rightarrow V \rightarrow W$	Causal	Mediator	Blocks the causal path
$U \leftarrow V \rightarrow W$	Noncausal	Confounding common cause	Blocks the noncausal path
$U \rightarrow V \leftarrow W$	None	Collider	Opens the noncausal path

The first of the three fundamental structures is causation or mediation. This construct is the most intuitive, and reflects a sequence in which a variable U plays a causal role in determining the value of V, which in turn plays a causal role in determining the value of W. For example, this construct would apply if taking a drug (U) caused a drop in blood pressure (V), which in turn increased the probability of fainting (W). In the causal structure, each of the variables is associated with each of the other variables, and this association is the result of a causal effect. However, if you were to condition on the variable V, the flow of information between U and W would be blocked. Continuing the example of a drug that causes a change in blood pressure, once you have controlled for a person's blood pressure, there is no longer any association between taking the drug and fainting.

The second fundamental structure is confounding. In this construct, a variable V is a common cause of two additional variables U and W. For example, parental socioeconomic status (V) might be a common cause of both enrollment in private school (U) and academic achievement (W). Much like the causal structure, each of the variables in the confounding structure is associated with each of the other variables. However, in the case of the confounding structure, the association between U and W is not causal. Rather, this association is induced by the common cause V. If you were to condition on V, the flow of information between U and W would be blocked. Returning to the example of private school enrollment and academic achievement, controlling for parental socioeconomic status blocks the noncausal association and thus would be useful for studying the causal effect of private school enrollment on academic achievement.

The third and final fundamental structure is the collider. In this construct, two variables U and W jointly play a causal role in determining the value of a third variable, V. A standard example is provided by Pearl (2009b): the weather (U) and a sprinkler (W) both play a role in determining whether the grass is wet (V). In the collider structure, both U and W are causally associated with V, but U and W are not associated with each other. However, conditioning on V would create association between U and W where previously no association existed. Returning to the sprinkler example, if you know that the grass is wet (conditioning on V), then knowing the value of either U or W can provide information about the other. For example, if the sprinklers were not used, then it must have rained. Because this induced association is not causal, it represents a form of bias called endogenous selection bias. See Elwert and Winship (2014) for a discussion and additional examples.

Each of the three fundamental structures that form a graphical model corresponds to a source of association and to a source of bias. Whether an association is biasing (that is, noncausal) depends not only on a set of causal assumptions (that is, the edges in a causal graph) but also on the conditioning decisions you make during an analysis. A detailed

understanding of the relationships between the structures in Table 1 and the sources of association and bias can help you reason about causal relationships among variables in your model. Once you have assembled these causal relationships into a causal graph, you can use the rules of causal graphs to assess the extent to which your causal model is consistent with the available data. This is discussed in the next subsection.

Statistical Properties of Causal Graphs

The ideas summarized in the preceding subsection are extended to paths that have an arbitrary length by the notion of *d-separation*. A path in a causal graph is said to be d-separated by a set of variables $\boldsymbol{X}$ if either of the following conditions holds:

- The path contains a chain $U \rightarrow V \rightarrow W$ or a fork $U \leftarrow V \rightarrow W$ such that $V \in \boldsymbol{X}$.
- The path contains a collider $U \rightarrow V \leftarrow W$ such that $V \notin \boldsymbol{X}$ and such that no descendant of V is in $\boldsymbol{X}$.

A path that is d-separated is said to be *blocked*. A path that is not d-separated is said to be *d-connected* or *nonblocked*. This terminology reflects the flow of information through the paths in a causal graph. If a path is blocked, then information does not flow through that path. If every path between two sets of variables is blocked, then any two variables (one from each set) are not associated. That is, the sets are independent. Thus, d-separation is a tool that converts a set of causal assumptions (the edges in a causal graph) into statements about statistical independence.

These statistical independence statements play two useful roles. First, you can use these statements to determine whether your causal model is consistent with the data generating process, or you can use them to distinguish between competing models on the basis of the independence statements that those models encode. For an example, see "Example 4: Distinguishing between Alternative Models." Second, you can make a careful choice of the set $\boldsymbol{X}$ to define an unbiased estimator for a causal estimand (for example, the average treatment effect). This is the goal of causal effect identification, as discussed in the next subsection. For an example, see "Example 2: Estimating a Causal Effect."

Causal Effect Identification

At this point, you have seen that once a causal graph has been specified, a set of graphical properties dictates all the causal and noncausal associations between the variables in the model. Moreover, these associations can be blocked or nonblocked as the result of conditioning on certain variables. This principle forms the basis for causal effect identification.

The goal of a causal analysis is to examine the association between one or more treatment variables and one or more outcome variables and to isolate the part of the total association that can be attributed to causal mechanisms. If it is possible to isolate and remove all noncausal association between treatment and outcome variables, leaving behind only causal association, then the causal effect is said to be *identified*. *Identification analysis* is the process of determining whether a causal effect can be identified and, if so, how to identify that effect.

Identification and Estimation by Adjustment

One intuitive approach to identification is to examine the sources of association between the treatment and outcome variables and determine which of these sources is causal. This is the basis for *identification by adjustment*, in which you attempt to find a set of variables that will block all noncausal paths between the treatment and outcome variables while leaving all causal paths unblocked. If such a set of variables exists, the causal effect is identified and the set of variables is called an *adjustment set*. For a set of treatment variables $\boldsymbol{T}$ and a set of outcome variables $\boldsymbol{Y}$, a set of observed variables $\boldsymbol{X}$ is a valid adjustment set if all the following conditions are present (Shpitser, VanderWeele, and Robins 2010; Perković et al. 2018):

- $\boldsymbol{X}$ blocks all noncausal paths between $\boldsymbol{T}$ and $\boldsymbol{Y}$.
- No variable in $\boldsymbol{X}$ lies on a causal path or descends from a causal path from $\boldsymbol{T}$ to $\boldsymbol{Y}$.
- No variable in $\boldsymbol{X}$ is a descendant of any variable on a causal path (except possibly the variables in $\boldsymbol{T}$).

You can use the CAUSALGRAPH procedure to find valid adjustment sets by specifying the METHOD=ADJUSTMENT option or the METHOD=BACKDOOR option. Once you have chosen a valid adjustment set $\boldsymbol{X}$, you can estimate the

total causal effect of the treatment $\boldsymbol{T}$ on the outcome $\boldsymbol{Y}$ by using the stratification estimator (Shpitser, VanderWeele, and Robins 2010; Elwert 2013). For discrete data, this estimator has the form

$$P(\boldsymbol{Y} = \boldsymbol{y}|\mathrm{do}(\boldsymbol{T} = \boldsymbol{t})) = \sum_{x} P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{X} = \boldsymbol{x})P(\boldsymbol{X} = \boldsymbol{x})$$

where the do-operator is intended to emphasize the interpretation of a causal effect as the result of an action or intervention (Pearl 2009b). This formula is exact in the asymptotic limit (so that each of the probabilities in the preceding expression are known exactly) and for countably many values of $\boldsymbol{X}$. However, in practice a causal effect must be computed from a finite sample. Moreover, the number of terms in the sum (the number of possible values $\boldsymbol{x}$) can be quite large. A typical approach to circumvent these difficulties is to replace the nonparametric terms in the stratification estimator by some parametric or semiparametric form that accounts for the biasing effects of the covariates in the set $\boldsymbol{X}$.

One approach that you can use to simplify the stratification estimator is to model the effect of the variables $\boldsymbol{X}$ on the treatment $\boldsymbol{T}$. For example, matching and weighting methods that are based on propensity scores are general methods of adjustment for the effect of covariates on the treatment. These methods are available in the PSMATCH and CAUSALTRT procedures. Another possible approach is to model the effect of the variables $\boldsymbol{X}$ on the outcome $\boldsymbol{Y}$ (more exactly, on the potential or counterfactual outcomes). This is the logic of regression adjustment methods. Such methods are available in the CAUSALTRT procedure. Finally, you can simultaneously model the effect of the covariates on both the treatment and the outcome variables. This leads to so-called "doubly robust methods." PROC CAUSALTRT also supports some doubly robust methods.

It is also possible to perform adjustment in the design stage of an observational study. For example, you can specify particular inclusion criteria so that only individuals with certain characteristics are included in the sample. You should also take care to avoid (or at least recognize) the possibility that a variable has been conditioned on as a result of survey nonresponse or loss to follow-up (Elwert 2013).

Identification and Estimation by an Instrumental Variable

An alternative approach to identification is to use an instrumental variable. This approach is particularly useful when using adjustment to identify a causal effect is not possible because there is unmeasured confounding between a treatment variable and an outcome variable. The instrumental variable method attempts to find a surrogate variable (the instrument) that is associated with the treatment variable but not associated with the outcome variable except through the treatment variable. If such an instrument exists, then the causal effect is identified. For single variables T, Y, and Z, and a set of variables $\boldsymbol{X}$, Z is an instrumental variable for the direct effect of T on Y if all the following conditions are met (Van der Zander, Textor, and Liśkiewicz 2015):

- There is a single causal path from T to Y that consists of a single edge.
- T and Z are d-connected conditional on $\boldsymbol{X}$.
- $\boldsymbol{X}$ d-separates Z and Y in G_c (the graph formed by taking the original DAG and removing the edge $T \rightarrow Y$).
- $\boldsymbol{X}$ consists of either ancestors of Y or ancestors of Z that are not descendants of Y (or both of these ancestors).

Specifically, this is the definition of an ancestral instrumental variable. In the CAUSALGRAPH procedure, you can specify the METHOD=IV option in the PROC CAUSALGRAPH statement to find ancestral instrumental variables. Once you have chosen a valid instrumental variable Z and its corresponding (possibly empty) conditioning set $\boldsymbol{X}$, you can use regression modeling to estimate the direct effect of T on Y.

EXAMPLE 2: ESTIMATING A CAUSAL EFFECT

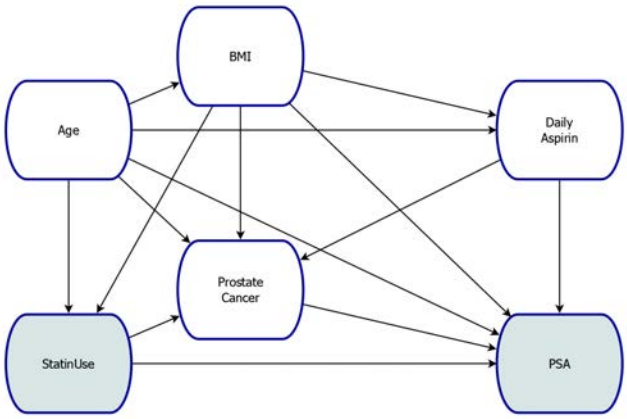
This example features a small causal model and a simulated data set. The purpose of this example is to demonstrate how to use the CAUSALGRAPH procedure in combination with other procedures to obtain an estimate of a causal effect that has a valid causal interpretation. Although this example is idealized in the sense that the true data generating process is known exactly, it nevertheless provides an important demonstration of the use of a causal model to identify and eliminate sources of bias in a causal analysis.

The causal model shown in Figure 6 is used in Ferro et al. (2015) to examine the relationship between the use of statin drugs and levels of prostate specific antigen (PSA). The model includes the following variables:

- **StatinUse**: the treatment variable
- **PSA**: the outcome variable
- **Age**: the subject's age in years
- **BMI**: the subject's body mass index in kilograms per square meter
- **ProstateCancer**: indicator for diagnosis with prostate cancer
- **DailyAspirin**: indicator for daily aspirin regimen

The treatment variable **StatinUse** is an indicator for whether the subject is taking any member of the statin class of drugs. The outcome variable **PSA** is the measured total circulating levels of PSA in a subject's blood, measured in ng/mL.

Figure 6 A Causal Model for the Effect of Statin Drug Use on PSA



The first 10 lines of a simulated data set are shown in Figure 7. The code to create this data set is available online.

Figure 7 First 10 Lines of the Simulated Data Set

Obs	Age	BMI	Aspirin	StatinUse	Cancer	PSA
1	67.8219	26.5741	0	1	0	4.19232
2	56.3722	26.4821	0	0	0	4.92659
3	57.8478	26.4788	0	0	1	6.97768
4	74.1587	26.6207	1	0	1	6.88858
5	69.4976	26.3082	0	0	0	5.89829
6	71.6984	26.0927	1	0	0	4.87417
7	61.0182	26.8058	0	0	0	6.14052
8	69.9975	27.0862	0	0	0	4.87205
9	56.6626	26.4457	0	0	0	4.55604
10	56.7639	24.1488	0	0	1	5.45989

Given the causal model (Figure 6) that describes the data generating process, you can use the CAUSALGRAPH procedure to analyze the identifiability of the causal effect of **StatinUse** on **PSA**. The following code uses the procedure to list the valid adjustment sets that you can use to identify this causal effect:

```
proc causalgraph;
   model "StatinUse Effect on PSA"
      BMI ==> Cancer StatinUse PSA Aspirin,
      StatinUse ==> Cancer PSA,
      Cancer ==> PSA,
      Age ==> BMI StatinUse Aspirin PSA Cancer,
      Aspirin ==> PSA Cancer;
   identify StatinUse ==> PSA;
run;
```

The initial output from the CAUSALGRAPH procedure is shown in Figure 8 and Figure 9. These figures contain, respectively, a summary table for the variables in the models and a summary table for the nodes and edges in each model. These outputs provide a useful check that your model syntax accurately reflects the causal models that you want to analyze.

Figure 8 Summary of Variables in the Models

Variables in Model		
	N	**Variables**
Measured	6	Age Aspirin BMI Cancer PSA StatinUse
Unmeasured	0	

Figure 9 Summary of the Models

Graphical Model Summary						
Model	**Nodes**	**Edges**	**Treatments**	**Outcomes**	**Measured**	**Unmeasured**
StatinUse Effect on PSA	6	14	1	1	6	0

The list of adjustment sets that the procedure produces is shown in Figure 10. Notice that the null set does not appear in this list. This means that the marginal association between **StatinUse** and **PSA** cannot be used to estimate a causal effect that has a valid causal interpretation. In fact, it is shown later that the marginal association between these two variables would produce a biased estimate of the causal effect. Rather, you must use an alternative estimation strategy, such as estimation by adjustment that uses one of the adjustment sets in Figure 10.

Figure 10 Possible Adjustment Sets for the Model in Figure 6

Covariate Adjustment Sets for StatinUse Effect on PSA						
Causal Effect of StatinUse on PSA						
			Covariates			
	Size	**Minimal**	**Age**	**Aspirin**	**BMI**	**Cancer**
1	2	Yes	*		*	
2	3	No	*	*	*	

Figure 10 shows two valid adjustment sets, and you can use either of these sets to obtain an estimate for the causal effect of **StatinUse** on **PSA**. For the sake of comparison, this example uses both adjustment sets to compute the causal effect.

The simulated data set in Figure 7 is constructed so that the true value of the average treatment effect (ATE) is –0.608. Intuitively, this means that in an ideal randomized controlled trial, subjects assigned to treatment would have total measured PSA that is 0.608 ng/mL lower on average than subjects assigned to control. (The formal definition of the average treatment effect is based on counterfactual outcomes and is beyond the scope of this paper. For details, see Hernán and Robins (2018).) Simulation code that computes the true value of the ATE is available online.

This example uses the CAUSALTRT procedure to compute the average treatment effect from the available data. The following code uses the variables **Age** and **BMI** as adjustments to compute the causal effect of **StatinUse** on **PSA**:

```
proc causaltrt data=StatinPSA;
   class StatinUse Cancer Aspirin;
   psmodel StatinUse(event='1') = Age BMI;
   model PSA;
run;
```

The CLASS statement indicates which variables in the data set are classification variables. The PSMODEL statement identifies **StatinUse** as the treatment variable and the value '1' as indicating a treatment event. The PSMODEL statement also requests that the variables **Age** and **BMI** be used to model the probability of receiving treatment (that is, the propensity score). It is this propensity score model that constitutes the covariate adjustment. Finally, the

MODEL statement specifies **PSA** as the outcome variable. With these options, the CAUSALTRT procedure uses the inverse probability weighting method with ratio adjustment to estimate the ATE. See Lunceford and Davidian (2004) for details.

Adjusting for **Age** and **BMI** produces an estimated ATE of –0.632, as shown in Figure 11. This value is within one standard error of the known true value of –0.608.

Figure 11 Causal Effect Estimated By Adjustment for **Age** and **BMI**

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	**1**	5.3265	0.0289	5.2698	5.3832	184.17	<.0001
POM	**0**	5.9585	0.0147	5.9297	5.9874	405.20	<.0001
ATE		-0.6320	0.0320	-0.6947	-0.5693	-19.76	<.0001

The same analysis can be repeated using the three variables **Age**, **Aspirin**, and **BMI** as an adjustment set:

```
proc causaltrt data=StatinPSA;
   class StatinUse Cancer Aspirin;
   psmodel StatinUse(event='1') = Age Aspirin BMI;
   model PSA;
run;
```

Adjusting for these three variables produces an estimated ATE of –0.6207, as shown in Figure 12. As before, the estimated value is within one standard deviation of the known true value.

Figure 12 Causal Effect Estimated By Adjustment for **Age**, **Aspirin**, and **BMI**

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	**1**	5.3349	0.0285	5.2791	5.3907	187.37	<.0001
POM	**0**	5.9556	0.0146	5.9271	5.9841	409.08	<.0001
ATE		-0.6207	0.0311	-0.6817	-0.5597	-19.93	<.0001

Note that the analysis results for the two adjustment sets are quite similar. This is expected because both adjustment sets are valid for the estimation of the causal effect of interest. Even so, the addition of **Aspirin** as an adjustment variable does lead to modest improvements in the accuracy and precision of the estimate because including additional variables is often helpful to better block a biasing path when a finite data sample and a parsimonious function of the adjustment variables (in this example, a propensity score model) are used. (This is true only if the adjustment set remains valid after the inclusion of the additional variables!) In practice, the choice of an adjustment set (among equally valid alternatives) must balance the need for accuracy and precision against the costs of data collection and computation.

Finally, note that adjustment is absolutely required in this example in order to obtain an unbiased estimate of the causal effect. To see this, consider the marginal association between **StatinUse** and **PSA**. The following code computes the difference in **PSA** between subjects who are and are not taking statins:

```
proc ttest data=StatinPSA;
   class StatinUse;
   var PSA;
run;
```

The results are shown in Figure 13. The marginal (that is, associational) effect of taking statins is 5.419 – 5.931 = –0.512, and the 95% confidence limit for the mean difference excludes the true value of the causal effect. These results provide a computational demonstration of the fact that, in this example, the marginal association is a biased estimate of the true causal effect.

Figure 13 Marginal Effect of **StatinUse** on **PSA**

StatinUse	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		5.9308	5.9019	5.9598	0.9139	0.8939	0.9349
1		5.4194	5.3670	5.4717	0.9155	0.8800	0.9541
Diff (1-2)	**Pooled**	0.5115	0.4517	0.5712	0.9143	0.8967	0.9326
Diff (1-2)	**Satterthwaite**	0.5115	0.4516	0.5713			

EXAMPLE 3: COMMON ADJUSTMENT SETS FOR MULTIPLE MODELS

This example demonstrates how you can use the CAUSALGRAPH procedure to examine the identifiability of a causal effect when there is some uncertainty about the exact structure of the data generating process. In the CAUSALGRAPH procedure, you can specify more than one causal model and attempt to find an adjustment set that can be used for all the models. If such an adjustment set exists, then the causal effect can be identified regardless of which model better represents the true data generating process.

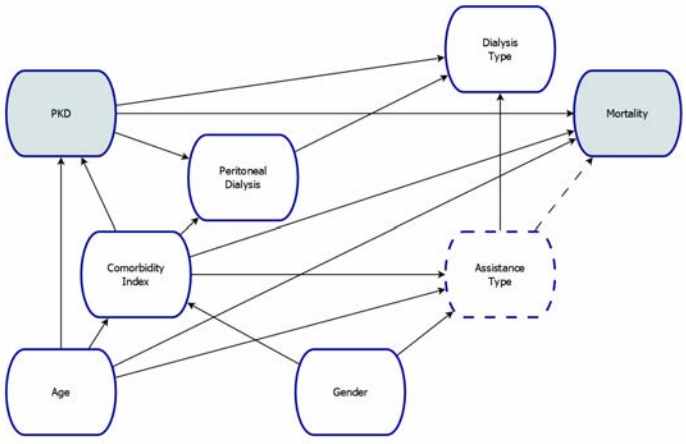
Two causal models are represented in Figure 14: one model includes all the nodes and edges in the figure and the other model is formed by deleting the dashed edge from **AssistanceType** to **Mortality**. These models were both formulated by Evans et al. (2012) to examine the effect on five-year mortality of polycystic kidney disease (compared to other nephropathies) among patients undergoing peritoneal dialysis. The models include the following variables:

- **PKD**: the treatment variable
- **Mortality**: the outcome variable
- **Age**: current patient age
- **Gender**: classification variable for patient gender
- **ComorbidityIndex**: summary index of patient comorbidities
- **PeritonealDialysis**: indicator for patients undergoing peritoneal dialysis
- **AssistanceType**: classification variable for types of medical assistance
- **DialysisType**: classification variable for type of peritoneal dialysis

The treatment (**PKD**) and outcome (**Mortality**) variables are shaded in Figure 14. As in Evans et al. (2012), it is assumed that research subjects receive peritoneal dialysis as a condition of enrollment in the research study. Thus, the variable **PeritonealDialysis** is conditioned on as a consequence of the design of the study. It is further assumed that any medical assistance received by a subject is not recorded, so the variable **AssistanceType** is unmeasured. This is indicated by the broken outline in Figure 14.

The dashed edge in Figure 14 indicates uncertainty about the structure of the data generating process, and thus about the form of the causal model. In this case, you are uncertain as to whether there is a causal effect of the type of medical assistance a subject receives (**AssistanceType**) on five-year mortality (**Mortality**).

Figure 14 Two Possible Causal Models for the Effect of PKD on Mortality



The following code uses the CAUSALGRAPH procedure to assess the identifiability of the causal effect in the models:

```
proc causalgraph common(only);
   model "Original Model"
      PKD ==> DialysisType Mortality PeritonealDialysis,
      Age ==> PKD ComorbidityIndex AssistanceType Mortality,
      ComorbidityIndex ==> PKD PeritonealDialysis Mortality AssistanceType,
      Gender ==> ComorbidityIndex AssistanceType,
      AssistanceType ==> DialysisType Mortality,
      PeritonealDialysis ==> DialysisType;
   model "Reduced Model"
      PKD ==> DialysisType Mortality PeritonealDialysis,
      Age ==> PKD ComorbidityIndex AssistanceType Mortality,
      ComorbidityIndex ==> PKD PeritonealDialysis Mortality AssistanceType,
      Gender ==> ComorbidityIndex AssistanceType,
      AssistanceType ==> DialysisType,
      PeritonealDialysis ==> DialysisType;
   identify PKD ==> Mortality | PeritonealDialysis;
   unmeasured AssistanceType;
run;
```

You specify each of the two models (one model with the dashed edge and one model without the dashed edge) in a separate MODEL statement. Each MODEL statement must begin with a quoted string that provides a unique name for the model. This example uses the labels “Original Model” and “Reduced Model.” You use the IDENTIFY statement to specify the treatment and outcome variables. In this example, you also use the IDENTIFY statement to specify that the variable **PeritonealDialysis** has already been conditioned on.

Because you are interested only in determining whether there exists a common adjustment set that can be used to identify the causal effect in either of the two causal models, you use the COMMON(ONLY) option in the CAUSALGRAPH statement. If you want to search for adjustment sets that are unique to each model in addition to searching for common adjustment sets, you can use the COMMON option rather than the COMMON(ONLY) option, as shown in Example 4.

The adjustment sets are shown in Figure 15. There are two adjustment sets that you can use to estimate the causal effect of **PKD** on **Mortality**. The existence of these adjustment sets means that the causal effect can be identified as long as either of the two models in Figure 15 (that is, with or without the dashed edge) accurately reflects the data generating process. Both adjustment sets contain the variables **Age** and **ComorbidityIndex** in addition to **PeritonealDialysis** (the latter having been conditioned on by virtue of the study design). The first adjustment set is minimal, which means that you cannot remove either **Age** or **ComorbidityIndex** from the adjustment set and still have a valid covariate adjustment. The second adjustment set also contains the variable **Gender**, but this adjustment set is not minimal because you could remove **Gender** from the set and still have a valid adjustment.

Figure 15 Common Adjustment Sets for the Models in Figure 14

Covariate Adjustment Sets Common to All Models

Causal Effect of PKD on Mortality

			Covariates				
	Size	Minimal	Age	ComorbidityIndex	DialysisType	Gender	PeritonealDialysis
1	3	Yes	*	*			*
2	4	No	*	*		*	*

EXAMPLE 4: DISTINGUISHING BETWEEN ALTERNATIVE MODELS

Just as in the previous example, this example demonstrates how you can use the CAUSALGRAPH procedure to examine the identifiability of a causal effect when you have multiple plausible causal models. Unlike the previous example, this example considers the case that you must choose which model (if either) to analyze because there is no common adjustment set. In this case, you can use the CAUSALGRAPH procedure to find observationally testable properties to help you decide between the competing models.

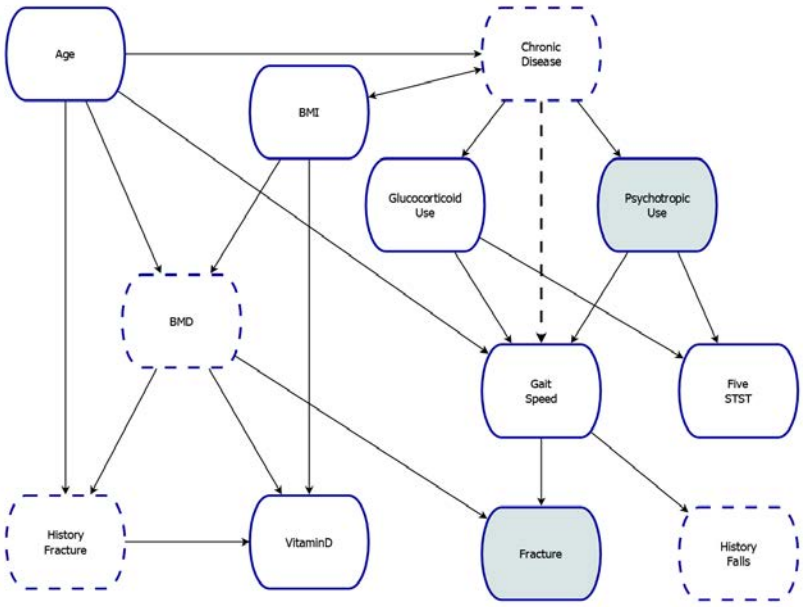
The two causal models shown in Figure 16 are adapted from Caillet et al. (2015) and examine the relationship between the use of psychotropic drugs and the risk of hip fracture in a care facility for the elderly. The models include the following variables:

- **PsychotropicUse**: the treatment variable
- **Fracture**: the outcome variable
- **Age**: age in years
- **BMD**: bone mineral density
- **BMI**: body mass index
- **ChronicDisease**: categorical variable that describes diagnosis with particular diseases
- **FiveSTST**: five-sit-to-stand-test score
- **GaitSpeed**: walking speed
- **HistoryFalls**: indicator for more than two falls in the previous six months
- **HistoryFracture**: indicator for previous hip fracture since age 55
- **GlucocorticoidUse**: indicator for currently taking glucocorticoids
- **VitaminD**: indicator for currently taking vitamin D

The treatment (**PsychotropicUse**) and outcome (**Fracture**) variables are shaded in Figure 16. For this example, it is assumed that the variables **BMD**, **ChronicDisease**, **HistoryFalls**, and **HistoryFracture** are not observed.

The dashed edge in Figure 16 indicates uncertainty over the structure of the data generating process. In this case, you are unsure whether **ChronicDisease** has a direct effect on **GaitSpeed** or whether the entire effect of **ChronicDisease** on **GaitSpeed** is mediated through **GlucocorticoidUse**.

Figure 16 Two Possible Causal Models for the Effect of Psychotropic Drug Use on Hip Fracture



As in the previous example, you can use the CAUSALGRAPH procedure to search for adjustment sets that can be used to identify the causal effect in both of the causal models:

```
proc causalgraph maxsize=min common imap;
   model "Base Model"
      Age ==> HistoryFracture BMD ChronicDisease GaitSpeed,
      BMI <==> ChronicDisease ==> PsychotropicUse GlucocorticoidUse,
      BMI ==> VitaminD BMD,
      PsychotropicUse ==> FiveTSTS GaitSpeed,
      GlucocorticoidUse ==> GaitSpeed FiveTSTS,
      BMD ==> HistoryFracture VitaminD Fracture,
      HistoryFracture ==> VitaminD,
```

```
      GaitSpeed ==> Fracture HistoryFalls;
   model "Modified Model"
      Age ==> HistoryFracture BMD ChronicDisease GaitSpeed,
      BMI <==> ChronicDisease ==> PsychotropicUse GlucocorticoidUse GaitSpeed,
      BMI ==> VitaminD BMD,
      PsychotropicUse ==> FiveTSTS GaitSpeed,
      GlucocorticoidUse ==> GaitSpeed FiveTSTS,
      BMD ==> HistoryFracture VitaminD Fracture,
      HistoryFracture ==> VitaminD,
      GaitSpeed ==> Fracture HistoryFalls;
   identify PsychotropicUse ==> Fracture;
   unmeasured ChronicDisease BMD HistoryFalls HistoryFracture;
run;
```

You specify the models, the causal effect, and the unmeasured variables just as in the previous example. In this example, you use the labels "Base Model" and "Modified Model" to identify the two possible causal models. You use the COMMON option in the CAUSALGRAPH statement to request that the CAUSALGRAPH procedure search for adjustment sets that are common to both models. The procedure also searches for adjustment sets that are unique to each model. To suppress the adjustment sets that are unique to each model, you can use the COMMON(ONLY) option as in the previous example. You use the MAXSIZE=MIN option to request only those adjustment sets that have the smallest possible number of variables. You use the IMAP option to request a set of independence properties that can be used to compare the two models. The IMAP option is discussed later in this example.

There is a single minimum-size adjustment set in the model "Base Model." This adjustment set, as shown in Figure 17, contains three variables: **Age**, **BMI**, and **GlucocorticoidUse**. In other words, if "Base Model" is the correct model, then it is sufficient to adjust for these three variables and it is not necessary to include any other variables in the adjustment set.

Figure 17 Minimal Adjustment Sets for "Base Model"

Covariate Adjustment Sets for Base Model

Causal Effect of PsychotropicUse on Fracture

			Covariates					
	Size	**Minimal**	**Age**	**BMI**	**FiveTSTS**	**GaitSpeed**	**GlucocorticoidUse**	**VitaminD**
1	3	Yes	*	*			*	

Unfortunately, there are no valid adjustment sets for the model "Modified Model." This means that, if "Modified Model" is the better description of the data generating process, then it is not possible to identify the causal effect by using an adjustment set. Because there are no adjustment sets for "Modified Model," there are correspondingly no adjustment sets common to both models. The nonexistence of an adjustment set for "Modified Model" or a common adjustment set is summarized in the following notes that PROC CAUSALGRAPH returns:

NOTE: There are no adjustment sets satisfying the specified criteria for Modified Model.

NOTE: There are no adjustment sets common to all models that satisfy the specified criteria.

Thus, of the two possible causal models for the effect of psychotropic drugs on the risk of hip fracture, the causal effect can be estimated using an adjustment set in one model but not in the other model. Thus, it would be very useful to determine which model is a better representation of the true data generating process. You can use the IMAP option to request the set of conditional independence statements that are encoded within each causal model. If such a conditional independence statement involves only observed quantities, then it can be tested in the available data to determine whether the independence predictions of the model are appropriate.

By default, the IMAP option produces a very concise list of independence properties called the *local Markov properties*. For each node in a DAG, the local Markov property states that the node is independent, conditional on its parents, of its nondescendants. The local Markov properties for "Base Model" are shown in Figure 18, and the local Markov properties for "Modified Model" are shown in Figure 19. These two sets of independence properties are identical except for row 7, where **ChronicDisease** has changed from a nondescendant (second column) of **GaitSpeed** for "Base Model" to a parent (third column) of **GaitSpeed** for "Modified Model." This change is expected because it reflects the single-edge difference between the two models.

Figure 18 Independence Properties for "Base Model"

	Local Conditional Independencies for Base Model	
	Independence Sets	**Conditioning Set**
1 Age	BMI	
2 BMD	ChronicDisease FiveTSTS GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	Age BMI
3 BMI	Age ChronicDisease FiveTSTS GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	
4 ChronicDisease	BMD BMI HistoryFracture VitaminD	Age
5 FiveTSTS	Age BMD BMI ChronicDisease Fracture GaitSpeed HistoryFalls HistoryFracture VitaminD	GlucocorticoidUse PsychotropicUse
6 Fracture	Age BMI ChronicDisease FiveTSTS GlucocorticoidUse HistoryFalls HistoryFracture PsychotropicUse VitaminD	BMD GaitSpeed
7 GaitSpeed	BMD BMI ChronicDisease FiveTSTS HistoryFracture VitaminD	Age GlucocorticoidUse PsychotropicUse
8 GlucocorticoidUse	Age BMD BMI HistoryFracture PsychotropicUse VitaminD	ChronicDisease
9 HistoryFalls	Age BMD BMI ChronicDisease FiveTSTS Fracture GlucocorticoidUse HistoryFracture PsychotropicUse VitaminD	GaitSpeed
10 HistoryFracture	BMI ChronicDisease FiveTSTS Fracture GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	Age BMD
11 PsychotropicUse	Age BMD BMI GlucocorticoidUse HistoryFracture VitaminD	ChronicDisease
12 VitaminD	Age ChronicDisease FiveTSTS Fracture GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	BMD BMI HistoryFracture

Figure 19 Independence Properties for "Modified Model"

	Local Conditional Independencies for Modified Model	
	Independence Sets	**Conditioning Set**
1 Age	BMI	
2 BMD	ChronicDisease FiveTSTS GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	Age BMI
3 BMI	Age ChronicDisease FiveTSTS GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	
4 ChronicDisease	BMD BMI HistoryFracture VitaminD	Age
5 FiveTSTS	Age BMD BMI ChronicDisease Fracture GaitSpeed HistoryFalls HistoryFracture VitaminD	GlucocorticoidUse PsychotropicUse
6 Fracture	Age BMI ChronicDisease FiveTSTS GlucocorticoidUse HistoryFalls HistoryFracture PsychotropicUse VitaminD	BMD GaitSpeed
7 GaitSpeed	BMD BMI FiveTSTS HistoryFracture VitaminD	Age ChronicDisease GlucocorticoidUse PsychotropicUse
8 GlucocorticoidUse	Age BMD BMI HistoryFracture PsychotropicUse VitaminD	ChronicDisease
9 HistoryFalls	Age BMD BMI ChronicDisease FiveTSTS Fracture GlucocorticoidUse HistoryFracture PsychotropicUse VitaminD	GaitSpeed
10 HistoryFracture	BMI ChronicDisease FiveTSTS Fracture GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	Age BMD
11 PsychotropicUse	Age BMD BMI GlucocorticoidUse HistoryFracture VitaminD	ChronicDisease
12 VitaminD	Age ChronicDisease FiveTSTS Fracture GaitSpeed GlucocorticoidUse HistoryFalls PsychotropicUse	BMD BMI HistoryFracture

Unfortunately, this one difference between the local Markov properties of the two models involves several unobserved variables, so it cannot be tested directly. One possible alternative is to use the graphoid axioms (Pearl and Verma 1987; Geiger and Pearl 1988) to derive additional independence properties that follow logically from the set of properties in the tables in Figure 18 and Figure 19. Then perhaps some of these derived properties might involve only observed quantities.

However, this is a very laborious process. In fact, approximately 17,000 independence conditions can be derived from each of the tables in Figure 18 and Figure 19! Fortunately, the d-separation property that was discussed in the

section "Theory of Causal Graph Analysis" provides a tool that can be used to decide whether any two variables can be made conditionally independent. In this view, d-separation is a *global Markov property*. In fact, you can use the graphoid axioms to show that the local and global Markov properties are equivalent for causal graphs (Koller and Friedman 2009). In the CAUSALGRAPH procedure, you request the set of all global Markov properties for each model by specifying the IMAP=GLOBAL option.

As noted earlier, there are approximately 17,000 global Markov conditions for each of the two models in this example. However, if you consider only those independence properties that consist exclusively of observed variables, this number is reduced to 150 for "Base Model" and to 144 for "Modified Model." Moreover, the 144 properties for "Modified Model" are all shared by "Base Model." The remaining six properties for "Base Model" are unique to that model. These six properties are shown in Figure 20. Code is available online to produce the global independence properties, select the properties that are observationally testable, and then compare these properties in the two models.

You can perform statistical tests on the available data to assess the extent to which these six conditional independence properties are present in the data. If all six properties seem to be present, then you have evidence that "Base Model" is a better model of the data generating process. Otherwise, you might consider eliminating "Base Model" in favor of "Modified Model." Of course, you might also consider examining the 144 properties that are shared between the two models to examine the extent to which other parts of the models agree with the available data.

Figure 20 Independence Properties Unique to "Base Model"

Obs	Model	Set1	Set2	CondSet
1	Base Model	BMI	GaitSpeed	Age FiveTSTS GlucocorticoidUse PsychotropicUse
2	Base Model	BMI	GaitSpeed	Age FiveTSTS GlucocorticoidUse PsychotropicUse VitaminD
3	Base Model	BMI	GaitSpeed	Age GlucocorticoidUse PsychotropicUse
4	Base Model	BMI	GaitSpeed	Age GlucocorticoidUse PsychotropicUse VitaminD
5	Base Model	GaitSpeed	VitaminD	Age FiveTSTS GlucocorticoidUse PsychotropicUse
6	Base Model	GaitSpeed	VitaminD	Age GlucocorticoidUse PsychotropicUse

CONCLUSION

Unbiased estimation of causal effects from observational data is an increasingly common task faced by data scientists and applied statisticians. Such an analysis requires assumptions about the underlying data generating process and research design. As described in this paper, causal graphs are a useful and concise tool that can help you articulate these assumptions and analyze patterns of association in your data. Once a causal graph has been articulated, you can use the rules of association and bias in a causal graph to assess the feasibility of estimating a causal effect.

The CAUSALGRAPH procedure enables you to analyze one or more causal graphs in order to determine whether it is possible to estimate a causal effect that has a valid causal interpretation. When a causal effect is identified, you can use the procedure to produce an estimation strategy for the causal effect. The CAUSALGRAPH procedure can find an adjustment set any time such a set exists. It can also identify instrumental variables that you can use to estimate a causal effect. It includes support for multiple treatment variables and multiple outcome variables, as well as support for unmeasured or latent variables. Whether or not a causal effect is identified, the CAUSALGRAPH procedure provides tools that you can use to examine the observationally testable implications of a causal model. You can use these implications to assess the extent to which a causal model accurately represents a data generating process, or you can use them to choose between competing causal models of a process.

The examples in this paper illustrate the use of the CAUSALGRAPH procedure to find adjustment sets for a single model and for multiple models. An example also shows how you can combine the CAUSALGRAPH procedure with the CAUSALTRT procedure to produce an estimate of a causal effect. Finally, this paper shows how you can find and use the observationally testable implications of a set of causal models in order to distinguish between those models.

REFERENCES

Caillet, P., Klemm, S., Ducher, M., Aussem, A., and Schott, A.-M. (2015). "Hip Fracture in the Elderly: A Re-analysis of the EPIDOS Study with Causal Bayesian Networks." *PLoS One* 10:e0120125. https://doi.org/10.1371/journal.pone.0120125.

Elwert, F. (2013). “Graphical Causal Models.” In *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan, 245–273. Dordrecht: Springer.

Elwert, F., and Winship, C. (2014). “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.” *Annual Review of Sociology* 40:31–53.

Evans, D., Chaix, B., Lobbedez, T., Verger, C., and Flahault, A. (2012). “Combining Directed Acyclic Graphs and the Change-in-Estimate Procedure as a Novel Approach to Adjustment-Variable Selection in Epidemiology.” *BMC Medical Research Methodology* 12:156.

Ferro, A., Pina, F., Severo, M., Dias, P., Botelho, F., and Lunet, N. (2015). “Use of Statins and Serum Levels of Prostate Specific Antigen.” *Acta Urológica Portuguesa* 32:71–77.

Geiger, D., and Pearl, J. (1988). “On the Logic of Causal Models.” In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, edited by R. D. Shacter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, 136–147. Amsterdam: North-Holland.

Hernán, M. A., and Robins, J. M. (2018). *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC. Forthcoming.

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lunceford, J. K., and Davidian, M. (2004). “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.” *Statistics in Medicine* 23:2937–2960.

Pearl, J. (2009a). “Causal Inference in Statistics: An Overview.” *Statistics Surveys* 3:96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Pearl, J. (2014). “Comment: Understanding Simpson’s Paradox.” *American Statistician* 68:8–13.

Pearl, J., and Verma, T. (1987). “The Logic of Representing Dependencies by Directed Graphs.” In *Proceedings of the Sixth National Conference on Artificial Intelligence*, 374–379. AAAI Press.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. (2018). “Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs.” *Journal of Machine Learning Research* 18:1–62.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). “On the Validity of Covariate Adjustment for Estimating Causal Effects.” In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, edited by P. Grünwald and P. Spirtes, 527–536. Corvallis, OR: AUAI Press.

Thornley, S., Marshall, R. J., Jackson, R., Gentles, D., Dalbeth, N., Crengle, S., Kerr, A., and Wells, S. (2013). “Is Serum Urate Causally Associated with Incident Cardiovascular Disease?” *Rheumatology* 52:135–142.

Van der Zander, B., Liśkiewicz, M., and Textor, J. (2014). “Constructing Separators and Adjustment Sets in Ancestral Graphs.” In *Proceedings of the Thirtieth Conference on Causal Inference: Learning and Prediction*, edited by N. L. Zhang and J. Tian, 11–24. Corvallis, OR: AUAI Press.

Van der Zander, B., Textor, J., and Liśkiewicz, M. (2015). “Efficiently Finding Conditional Instruments for Causal Inference.” In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, edited by Q. Yang and M. Wooldridge, 3242–3249. Palo Alto, CA: AAAI Press.

ACKNOWLEDGMENTS

The author is grateful to Yiu-Fai Yung and Michael Lamm of the Advanced Analytics Division at SAS for their valuable assistance in the preparation of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Clay Thompson
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Phone: 919-555-1212
Email: clay.thompson@sas.com