**Chapter 20**

**Causal Networks**

# [H1] Learning Objectives

* Construct causal networks through repeated regression equations
* Predict probability of events in a network model
* Calculate the causal impact of treatment on an outcome using a network model
* Remove confounding in electronic health record data

# [H1] Key Concepts

* Network models
* Direct cause
* Ordered or sequenced variables
* Do operation
* Back-door path
* Markov blanket
* Directional separation
* Parents
* Children
* Co-parents

# [H1] Chapter at a Glance

Throughout this book, with exception of the Chapter 19 on association networks, we have assumed that there is one outcome. We would like to predict it, or relate it to several independent variables. Real life is, of course, more complex than this. In real life, there are multiple outcomes, and the outcomes themselves are interrelated. For example, survival is one outcome of care; cost is another; these two outcomes are inter-related—dead patients have no cost. In addition, some independent variables used to predict one outcome may be used to predict another outcome. Age can predict both survival and cost of care. In fact, some outcomes are predictors of subsequent variables. For example, age can predict hospital cost as well as later nursing home costs. Regression analysis cannot adequately answer these complexities. Analysis of multiple outcomes requires network modeling.

One could think of network models as multiple simultaneous regression equations, so in a sense, network models are an extension of regression, a topic we have repeatedly addressed in Chapters 11 through 15 in this book. In this chapter, we describe causal networks. We show how these networks are specified using regression analysis; how these networks measure the causal impact of treatment; and how causal networks could help reduce massive, high-dimensional data to a more manageable size.

**[H1] Causal Thinking Is Fundamental**

Without causation, there is no way of making sense of data in electronic health records (EHRs). In these records, thousands of variables are available. In such a high-dimensional observational data, there are a great deal of confounding. Even if we discover a relationship between two variables, it is not clear whether the relationship is an artifact of a common cause, a false association, or a true relationship.

**[H1] Use of Network Models**

In recent years, there have been increasing application of causal networks. Causal networks have been applied to understanding the relationship among comorbidities and whether one disease is a complication of another (Bang et al. 2016). The classification of diseases into comorbidities and complications is essential for analysis of data in electronic health records, as comorbidities are stratified and complications are not. If one is evaluating the impact of treatment in observational data, comorbidities can affect both treatment and outcome and must be statistically controlled. At the same time, complications are on the causal path of treatment to outcome and statistically controlling them would distort the relationship between treatment and outcome.

Another example of application of causal networks is given by studies of pharmacovigilance. Health authorities and courts may use causal networks to examine if a medication has caused adverse reaction. French pharmacovigilance center, for example, does so (Théophile et al. 2010; Rodrigues et al 2018).

Application to managerial issues is still less common. Managers may be interested in understanding which employee is responsible for a patient’s satisfaction ratings. A network model can separate out the contribution of team member to overall performance of the team. Financial officers may wish to understand the advantages of contracting with one or another nursing home in the context of bundled prices. Network models can separate out the effect of different organization to cost effectiveness of bundled prices. These kinds of application was described in Chapter 17.

Health insurers may wish to limit visits to specialists as a way to reduce cost and encourage in-plan use of primary care providers. Investigators have shown that narrow provider networks may reduce cost while having other implications on patient satisfaction or long term health (Atwood and Sasso 2016).

Kheirbek and colleagues (2015) describe a network model of causes of excessive patient boarding times in emergency rooms. They found that causes of delay had little to do with efficiency of the emergency room and that the more important root causes were back up in imaging and hospital bed availability.

In this book we have repeatedly tried to address causal analysis. We discussed causal control charts in Chapter 10. In Chapter 13 on propensity scoring, we discovered how to use regression to remove confounding in observational data. Removing confounding is a fundamental requirement of causal analysis. In Chapter 14 on multilevel regression, we have seen how patient characteristics can be controlled while examining the predictors of the performance of hospitals and practices. In Chapter 16 on covariate balancing, we have shown how stratification can help remove confounding. All of these chapters show procedures for removing confounding, a fundamental step for causal analysis. We have taken these steps without a comprehensive theory of what is causality and how it should be modeled. Now, in the final chapter of the book, we take on a more comprehensive look at causal modeling.

**[H1] Correlation Is Not Causation. So What Is Causation?**

We often hear that correlation is not causation. We agree. We know what causal analysis is not, but do we know what it is? There are four principles of causation that distinguish causal analysis from analysis of associations:

1. Causes and effects are associated with each other and therefor analysis should examine association among variables. Nothing new here. Any regression measures association between variables.
2. Causes occur prior to effects. In causal analysis the sequence of occurrence of variables matters. In regression, for example, the timing of occurrence of the events does not matter. Causal analysis must take advantage of the sequence information. Reverse association between variables are ignored in causal analysis. For example, mortality cannot cause a disease, although the two can be highly correlated.
3. There is a hypothesized mechanism for how the cause leads to the effect. Causal analysis requires description of a mechanism of the causation. This can be done in the statistical analysis through mediation analysis establishing that a variable describes the mechanism of the effect of cause on effect. It can also be done outside the numerical analysis by hypothesizing possible physical ways in which causes might lead to an effect. Causal analysis cannot be done if there is no explanation of the mechanism of the effect.
4. In the absence of the cause, the effect should not occur. Causal analysis differs from association analysis because it posits that a change in the cause will always affect a change in the outcomes. For example, the veteran administration has a model for predicting risk factors for suicide among veterans. This is based on association of various events with suicide. Unfortunately, removal of a risk factor for suicide, even though it is highly associated with suicide, does not change the probability of suicide. In contrast, if risk of suicide was based on causal analysis, removing it, would in fact reduce the probability of suicide. In this sense, causal analysis is about doing things and not just analyzing data. It is about measuring what will happen if we intervene in a situation.

The fourth principle of causal analysis is a crucial principle. It is referred to as *counterfactual*. Usually, the data show what happens when the cause is present. For example, the data show what happens to the patient who took the prescribed medication. That information tells us nothing about what would have happened to the same patient if he had not taken the medication. Of course, there are many patients for whom the cause is absent (i.e., who did not take the medication), but these patients differ from the patient who took the medication. To calculate causal impact, the analysis must simulate what would have happened if the cause were absent. This impact cannot be observed—it is not factual—which is why this principle is counterfactual.

# [H1] Comprehensive Analysis Is Necessary

In determining the cause of an effect, it is important to rule out alternative causes as possible explanations. This requires causal analysis to measure and analyze any possible alternative explanation of the effect. There are two problems with a plan to include all possible causes. First, the analyst may not know a complete list of alternative explanations. He could, of course, hypothesize some, but given the massive data in EHRs, many causes exist of which he will be unaware. Given millions of patients, he may not know all the unusual events that may have caused the observed data. His clinical experience is far more limited than the data in an EHR; therefore, he may not have been exposed to all possible causes.

The second problem is that a call to include all possible causes is a call to be comprehensive. Doing so leads to high-dimensional problems, at which point many normal statistical methods fall apart. For example, if we want to understand the causal impact of lung cancer surgery on survival, we must include the patients’ comorbidities in totality. In EHRs, the comorbidities could be thousands of different diagnoses. These include other cancer diagnoses, which obviously affect mortality from lung cancer. A cancer patient may also die from heart failure. Depressed cancer patients have worse outcomes than patients who are not depressed. The list goes on and on. Almost any disease may alter survival from lung cancer surgery. To analyze the causal impact of lung cancer surgery, given a diverse set of patients, one is forced to consider thousands of other comorbidities. Thus, real causal analysis relies on high-dimensional data.

# [H1] Key Concepts in Causal Networks

A causal network is a collection of interrelated causes and their effects. *Interrelated* means that one cause can have an effect on another cause. Effects can cause changes in each other. Almost any relationship between two variables is possible except the reverse causation. An effect cannot change its causes. A causal network is a collection of nodes and directed links among pairs of nodes. Each node represents one variable, and each link a relationship between a pair of variables.

Many of the concepts in multiple regression or multivariate analysis have equivalent but different terms in network analysis (see exhibit 20.1). In both network analysis and multivariate statistical analysis, outcome or response refers to a variable predicted from other, sometimes called independent, variables. In both, treatment refers to manipulation of the world to change outcomes – something one does in the real world and not just in data. Statisticians call a variable covariate when it affects both outcome and treatment. Network analysts refer to covariates as variables on the back-door path from outcome to treatment. Back-door is a set of overlapping pairs of associated variables, which we will define further later in this chapter. In network analysis, parent and children refer to direct causes and direct effects of a variable. Statisticians refer to controlling the effect of a variable as stratification. Network analyst call this conditioning. Exhibit 20.1 shows how multivariate and network analysts use different words and terminology to refer to the same concepts. This proliferation of different terminology is unfortunate as it reduces communication among the different branches of science.

**Exhibit 20.1** Comparison of Terminology in Network and Multivariate Analysis

|  |  |  |
| --- | --- | --- |
| **Term** | **Network Definition** | **Multivariate Definition** |
| Stratification | To conduct separate analysis for each level of a variable. In networking terminology, stratification is also referred to as *blocking a path* or *conditioning on a set of variables*. | |
| Outcome | A response variable that is measured after all other variables in the study. | |
| Treatment | A variable that reflects manipulations undertaken to affect outcome. | |
| Covariate | A node that, by itself or through other nodes, is associated with treatment and outcome. | A variable associated with both treatment and outcome. |
| A-cyclical | A network for which it is not possible to start from a node, follow the directed arcs, and return to the same node. | No variable is allowed to be both dependent and independent. |
| Parent | A node that has a directed arc to another node. Parents have a statistically significant association with effects independent of other variables. | A variable that occurs before another; the association between the two variables does not disappear in any subset of data. |
| Children | A node that receives a directed arc from another. A variable has a statistically significant relationship to its children independent of other variables. | A variable that occurs after another; the association between the two variables does not disappear in any subset of data. |
| Descendants of treatment | Any node reached from treatment following the arcs in the network. | Variables that are early or late treatment effects. |
| Collider or co-parent or common effect | A common effect of two causes. A node that, if stratified, would make two conditionally independent nodes dependent. | Two variables that have a statistically significant relationship with each other when a third variable is present but not when the third variable is absent. |
| Path | A set of nodes connected by arcs independent of the causal direction. | A set of overlapping pairs of associated variables. |
| Back-door path from outcome to treatment | A path from outcome to treatment without following the direction of the arcs and ending with an arc into treatment. | A path that starts from outcome and ends with treatment without including events that occur after treatment. |
| Blocked back doors | Stratification of a non-collider node on all back-door paths from outcome to treatment. | Stratification that removes association of all covariates with either treatment or outcome. |
| Markov blanket (d-separation) | Parents, children, and co-parents (other parents of children) of treatment. | Smallest set of stratified variables that would make a variable independent from all other variables. |

## [H2] Directed Arcs

A causal network is a collection of interrelated causes and their effects. *Interrelated* means that variables have effects on each other, sometimes in a chain, other times a common cause affecting several variables and still other times multiple causes having the same effect. If one variable directly causes another, then an arc is drawn between the two nodes. Unlike an association network, causal networks use directed arcs. The direction of the arc is from cause to effect. The causal impact of fatigue on medication error is shown in exhibit 20.2.

**Exhibit 20.2** Causal Impact of Fatigue on Medication Errors

Exhibit 20.2 shows two variables, medication error and long hospital stays, each in a node. This display shows that medication errors lead to prolonged hospital stays. The link between the nodes shows that these two variables are associated with each other. The arrow in the link shows that medication error causes long hospital stays, not vice versa. If there was no connection between the two variables, they would be considered independent.

**[H2] Not a Cyclical Graph**

To make sure that effects do not change causes, the network is assumed to be a-cyclical, meaning you cannot start from any variable, follow the paths in the network and end up the same place. All causal networks are, by definition, directed a-cyclical graphs, or DAGs for short. Every arc is directed; there are no cycles in the network. Causal networks are not suitable for analysis of cyclical causes. This is not to say that in real life there are no cycles of causal effects. However, causal networks cannot be used to study circular causations. In addition, causal networks cannot be used to study partially directed networks. The methods of causal networks assume that we are dealing with a complete DAG. Exhibit 20.3 shows a graph that is not cyclical, you cannot start from any node in this graph and cycle back to the same node.

**Exhibit 20.3** Causal Network for Long Hospital Stays

**[H2] Only Direct Causes Are Shown**

Causal networks show only the direct causes of the effects. Indirect causes can be calculated from direct effects but are not displayed in the network. If two variables are indirectly related to each other, the viewer can start from one variable, follow the directed links, and reach the other. If two variables are unrelated, one cannot follow the links shown in the network to reach from one to the other. In exhibit 20.3, we do not see a direct causal impact between provider fatigue and long hospital stay. This does not mean that provider fatigue does not affect long hospital stay. It does. It affects long hospital stays through medication errors. In a network model you can follow the directed arcs to trace the downstream causal effects of a variable. Only the direct causes are shown and the rest are inferred.

Furthermore, all direct causes are shown. Exhibit 20.3 shows two competing causes of long hospital stays. Patients may stay longer in the hospital because they have had a medication error or they are sicker than average hospitalized patient. A causal model should show all causes of the effect, therefore both severity of patients’ illness and medication errors must have a direct causal impact on long hospital stays.

**[H2] What Is Not Shown Has Meaning**

In a network model, what is not shown has meaning and implies lack of direct causal relationship. If it is not possible to follow the arcs and reach a node from another, then the two are independent. A typical network shows much more independence than it shows causal relationships. In exhibit 20.3, we see four causal relationships depicted by directed arcs and several situations for which a directed arc could have been present but is not:

1. We do not see a link between severity of the patient’s illness and provider fatigue. There is no way to start from provider fatigue and arrive at severity of the patient’s illness. This means the two concepts are independent from each other.
2. We do not see a direct causal link between long hospital stay and provider fatigue. Long hospital stay is an effect of provider fatigue. To add a directed arc would have created a cycle in the network, it would have shown an effect changing the cause. Where there are directed links, we do not see reverse links creating circular causation.

**[H2] Causal Chain**

A causal chain refers to the situation where one cause affects the cause of another event. In exhibit 20.3, we see causal chains. The provider fatigue has an impact on the length of stay, but this impact is mediated through medication error. Provider fatigue is shown to cause medication errors, and medication errors are shown to cause long hospital stays. These three variables are said to be in a *causal chain*.

**[H2] Common Effect**

Common effect refer to the situation where multiple causes have the same effect. In exhibit 20.3, we also see a common effect. A common effect occurs when more than one cause lead to the same effect. Severe illness and provider fatigue both cause medication errors.

**[H2] Common Cause**

A common cause refers to the situation where one cause leads to multiple effects. In exhibit 20.3, we also see a common cause. A common cause is when the same cause has multiple effects. Severe illness causes both medication error and long hospital stays.

[H2]Irrelevant Variables Are Not Shown

To make networks easier to understand, a variable that is not related to any other variable in the network is not shown. So in Exhibit 20.3, there is no node that is not connected to the entire network in at least one place. If there were such a node, then that variable is independent from all variables in the network, so irrelevant to our analysis.

## [H2] Genealogy in Networks

Family genealogy can be used to describe particular relationships in a causal network. One can refer to direct cause of a variable as its *parent*. Exhibit 20.3 shows “provider fatigue” is parent to medication error. Provider fatigue is not parent to long hospital stays, as it is not a direct cause of long hospital stays. Direct effects of causes are referred to as *children* of the variable. Returning to the same exhibit, “long hospital stay” is a child of “medication error.” Parents of children of a variable are referred to as co-parents. If there were no link between “severe illness” and “medication error,” then severe illness would be considered a co-parent of medication error. Indirect causes of a variable are referred to as *ancestors* of the variable. “Medication error” is parent to “long hospital stay,” and “provider fatigue” is an ancestor of “long hospital stay.” Indirect effects of a variable are referred to *descendants* of the variable. In exhibit 20.3, “medication error” is a child, and “long hospital stay” is a descendent of “provider fatigue.” Note that if we display parents of variables, then the entire network of children and co-parents can be easily read from the display.

## [H2] Removing Spurious Correlations

In contrast to an association network, causal networks reduce spurious correlations. If the correlation between two variables results from other variables in the network, no link between the two variables is shown, even if the correlation may be statistically significant. For example, the correlation between two causes appears and disappears depending on whether their joint effect is stratified. In one subset of the data, where the effect is present, the correlation exists; in another, it does not. Network models ignore these correlations on grounds that they are function of causes already modelled. They are spurious in the sense that the correlation is an artifact of stratifying common effects.

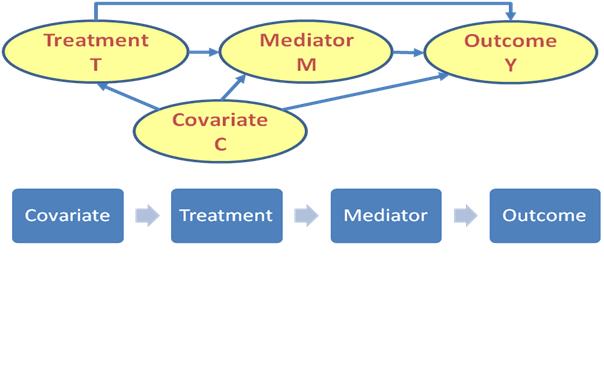
Network models also do not show correlations resulting from a common cause. For example, aging may lead to Alzheimer’s disease, and statin medications are taken by older patients. Therefore, a correlation may exist between these two variables—they are both affected by aging. The correlation between these two variables may disappear if we stratify aging and look at the relationship at different age levels. If stratifying aging removes the correlation, then causal networks do not show a relationship between Alzheimer’s and statin. If the correlation persists, the relationship is included. Causal analysis can remove spurious correlations.

# [H1] Relationship between Regression and Causal Networks

A standard regression can be expressed, under very general assumptions, as a network model. Consider, for example the following regression equation:

In this equation, *Y* is the outcome; is the intercept; is set of covariates that occur prior to treatment (e.g., patient’s comorbidities); is an estimated impact of the covariates on the outcome; is a series of variables that occur after treatment and mediate the impact of treatment on outcome (e.g., complications); is the parameter estimated for the impact of medicating variables on the outcome; *T* is treatment; is the impact of treatment on the outcome; and indicates a standard, normally distributed error term. This regression equation could be easily represented as the network in exhibit 20.4.

**Exhibit 20.4** Network Representation of



Note that regression analysis is a study of association among the variables; therefore, the network representation of a regression equation should be an association network. In exhibit 20.4, we have assumed a causal impact among the variables, including a particular sequence among the variables. The information on sequence is not in the regression equation but can be assumed if we choose our variables carefully.

If we want a causal model of regression, we must assume that outcome occurs last. Covariates (e.g., comorbidities, medical history) occur before treatment and can affect mediators and the outcome. Mediating events (e.g., complications) occur after treatment and before the outcome. Statisticians prohibit the use of mediating factors in regression equations, as these variables distort the relationship between treatment and outcome.

In addition, we are assuming that the independent variables in regression have a clear mechanism for affecting the outcome. These assumptions cover sequence and mechanism. The association and counterfactual assumptions can be verified empirically.

While regression and causal networks have a great deal of similarities, there are also large differences. First, in causal networks, the parameters are measured independent of other variables. All confounding is removed before measuring the impact of a variable on another. In contrast, regression analysis does not actively remove confounding. Second, in regression, intercorrelations among the variables force some variables to remain in regression and other variables to drop. In contrast, network models include redundant variables.

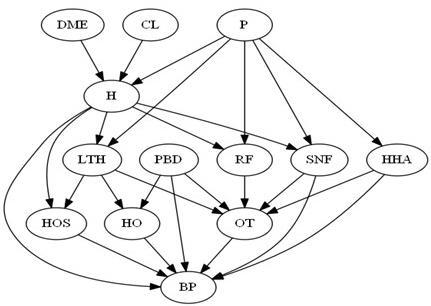
Sometimes, the variables are intentionally dropped from the analysis. For example, we already mentioned that mediating variables can distort the relationship between treatment and outcome and that statisticians intentionally drop these types of variables from regression. In network analysis, mediating variables are kept in the network. In predicting survival from bypass surgery, complications must be dropped from a regression analysis. Including complications in the regression will attribute most mortalities to complications and therefore underestimate mortalities resulting from bypass surgery. In a network model of the same set of facts, it is OK to include complications in the network; although they are still excluded when measuring causal impact of treatment on outcome.

Regression selects an optimal set of variables and explicitly drops all variables in the causal path of treatment to outcome; network models do not. In other words, regression selects a small set of variables while network models include all variables. For the entire sample, regression equations are highly accurate because optimal parameters have been selected. These equations may be less accurate in a specific subset of data or when one of the selected variables is missing. What proved optimal for the entire data may not be accurate in a subset of data. In contrast, causal analysis includes redundant variables and therefore may be robust to missing values. Because the parameters are estimated independent of other variables, the effect of the variables in a different subset of data and the entire sample may be consistent.

# [H2] Causal Networks as Multiple Regressions

Causal networks can be represented as several interrelated regressions. To demonstrate how a network can be broken down into several regression models, we can examine a relatively complex network. The network in exhibit 20.5 describes how various cost overruns may occur in a 90-day episode of treatment for hip fracture. The Center for Medicare and Medicaid Services provides a bundled payment to the hospital, which puts the hospital at risk if the cost exceeds the bundled amount. In this network, durable medical equipment cost (DME), clinical laboratory tests (CL), and physician bills (P) are assumed to affect hospital (H) costs. Likewise, hospital costs are assumed to affect long-term hospital cost (LTH), rehabilitation facility cost (RF), skilled nursing facility cost (SNF), hospice cost (HOS), and eventually CMS’s bundled payment cost (BP). This follows from the assumption that sicker patients will have higher costs throughout different institutions.

**Exhibit 20.5** Predicting Bundled Payment Cost Overrun from Related Costs



*Note*: DME = Durable medical equipment, CL = clinical laboratory, P = physician, H = hospital, LTH = long-term hospital, RF = rehabilitation facility, SNF= skilled nursing facility, HHA = home health agency, PBD = Part B drug, HO = hospital outpatient, OT = outpatient therapy, HOS = hospice, BP = bundled payment cost overruns.

Physician billing is expected to affect home health agency cost (HHA), SNF, RF, and LTH. LTH and Part B drug cost (PBD) are expected to affect hospital outpatient (HO) cost, and PBD is expected to affect HO and outpatient therapy cost (OT). OT is also affected by RF, SNF, and HHA. BP is affected by seven variables, including H, HOS, HO, OT, PBD, SNF, and HHA. Exhibit 20.5 is a relatively complex network structure, with many relationships, including a large number of assumptions about independence of variables. This network can be shown as a number of interrelated regressions. For every node, the parents to the node are predictors of the node; all other nodes are either blocked by the parents and therefore would not matter if they are included in the regression or are descendants of the variable and should not be included in the regression. Thus, the cost at discharge from the hospital is a function of the parents to hospital node, represented in mathematical terms as the equation

.

Similarly, the regression equation for the HO variable can be written as a regression on parents of HO:

.

Note that some variables are listed as independent variables in multiple regression equations. Also, dependent variables in some regression equations become independent variables for later regression equations. For example, the regression equation for BP can be written as a function of the seven parents of BP:

.

These parents were response variables in previous regressions. In this fashion, a network model can be broken into separate but interrelated regressions.

# [H1] Predicting Probability of an Event

To predict from a network, analysts can use both causal and association links. A causal prediction moves along a time dimension, from the past to the future; from cause to effect. A prediction based on association can go against time—against the arcs in the network—to predict something in the past. Thus, Alemi and colleagues (2018) predicted that a patient who has had multiple surgeries (with associated opioid medications for pain) may in time have a prescription abuse problem, which would be a causal prediction. Or one can predict from consequences of opioid addiction (e.g., self-injury) that the patient may have already abused prescribed medications. The former is a prediction in time and the latter is a detection of a missed clinical problem. Association links can be used to assess the probability of past or future events. Causal links can only be used to predict future events. Movement along the directed arcs inside a causal network is always a move forward in time. Similar to hanging mobiles, a change in one part of a network can reverberate everywhere in the network. The change follows both the causal pathways and the associations to affect the probability in all remaining parts of the network.

Exhibit 20.6 shows a network that relates a patient’s severity of illness to a clinician’s choice of treatment, then to an outcome. In this network, severity is a parent to patient preferences on resuscitation, treatment choice, and outcome. The parents of the treatment node are severity, resuscitation, and provider’s decision. The parents of the outcome are the treatment received and the severity of the illness. These parent–child relationships indicate a dependency in the data. Furthermore, these interdependencies allow us to estimate the joint probability distribution without having to look at all possible combinations of variables.

**Exhibit 20.6** Network Model of Treatment

**S: Severity of Illness**

**R: Do Not Resuscitate**

**T: Treatment**

**O: Outcome**

**M: Provider’s Decision**

Let us write the equation for estimating the joint probability of the events in the network in exhibit 20.6. To save space, we use the abbreviation of each variable: O for outcome, S for severity, T for treatment, M for physician’s choice of treatment, and R for patient’s preferences not to be resuscitated. Let us start at the outcome. This is called an end node, as it has no children. The equation for calculating the probability of outcome is given conditional to the probabilities of its parents (nonparents are not relevant):

.

We can also calculate the probability of treatment from the product of conditional probabilities of treatment given its parents, as

.

Finally, we calculate the probability of severity, do not resuscitate, and physician’s choice of treatment. These events have no parents, and their probabilities are merely marginal probabilities: p(S), p(R), p(M). Now we can put all five calculated probabilities together to estimate the joint probability of all events in the tree. Note how a variable shows up as a predicted item in one equation and a condition in one or more equations. The assumption that we need to only condition on parents has radically simplified what conditional probabilities are needed. Furthermore, all needed terms are either directly available or calculated from prior equations, allowing us to calculate the joint probability of any event.

**Exhibit 20.7** Probability Distribution for Events in a Network

|  |
| --- |
|  |

Once we have the joint distribution of the events, we can use it to calculate conditional probabilities of events given that other events have already occurred. The conditional probability is the probability of the event divided by the probability of all possible events. So for the probability of a particular outcome for a patient who is severely ill, we divide the joint probability of outcomes for severe patients by the probability of observing severe patients. Think of it this way—we have selected all patients who are severely ill, and within these patients, we look at the frequency various outcomes. Consider the following calculation:

Note that in this calculation, the joint distribution of two variables is needed—outcome and severity. Earlier, we had calculated the joint distribution of all five variables. To move from the joint distribution of all five variables to fewer variables, we have to sum out the missing variables and calculate marginal tables. In this case, the three variables (treatment, physician decision, and resuscitation preferences) are missing in the joint distribution; they are summed out of the joint distribution of all five variables.

Let us look at another set of calculations. Suppose we want to know the probability of having a positive outcome among treated, though severely ill, patients. If the joint distribution is known, this conditional probability could be easily calculated. As before, in this calculation, the joint distribution of three variables (outcome, severity, and treatment) is needed. We had previously calculated the joint distribution of all five variables. To move from the joint distribution of all five variables to fewer variables, we have to sum out the missing variables. Then, we can calculate the probability of the outcome as

## [H1] A Numerical Example

To see how the probability of an event is calculated in a network, we provide a numerical example. Let us start with a simple tree with the four variables: patient’s severity, patient’s preferences for a “do not resuscitate” order (DNR), treatment choice, and outcome. Exhibit 20.8 provides the distribution of the variables. Some variables, such as severity of illness (S) and presence of a DNR, are marginal probabilities calculated as percent of patients in the sample that have the condition. Other variables, such as treatment and outcome, are conditional probabilities, and are calculated from the sample by conditioning on parents of the variables. The effect of changes in a network can be calculated by repeatedly going through two steps known as *joining* and *eliminating*. The joining works in the same manner as an inner join in SQL tables. Elimination is done by summing over the variable that we are not interested in. An example can demonstrate how these steps are carried out.

**Exhibit 20.8** Repeated Joins and Elimination Steps

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **p(S)** | | **Joins with** | **p(DNR | S)** | | | | **Equals =** | **p(DNR , S)** | | |  | | |
| **S** | **P** | **S** | **DNR** | **P** | | **S** | **DNR** | **p** |
| Yes | 0.40 | Yes | Yes | 0.80 | | Yes | Yes | 0.32 |
| No | 0.60 | Yes | No | 0.20 | | Yes | No | 0.08 |
|  | | No | Yes | 0.10 | | No | Yes | 0.06 |
| No | No | 0.90 | | No | No | 0.54 |
|  | | | | | | | | | | | | | |
| **p(DNR , S)** | | | **Joins with** | **p(Tx|DNR, S)** | | | | | **Equals =** | **p(Tx, DNR, S)** | | | |
| **S** | **DNR** | **p** | **DNR** | **S** | **Tx** | | **p** | **DNR** | **S** | **Tx** | **p** |
| Yes | Yes | 0.3 | Yes | Yes | Yes | | 0.10 | Yes | Yes | Yes | 0.03 |
| Yes | No | 0.1 | Yes | Yes | No | | 0.20 | Yes | Yes | No | 0.06 |
| No | Yes | 0.1 | Yes | No | Yes | | 0.00 | Yes | No | Yes | 0.00 |
| No | No | 0.5 | Yes | No | No | | 0.00 | Yes | No | No | 0.00 |
|  | | | No | Yes | Yes | | 0.30 | No | Yes | Yes | 0.02 |
| No | Yes | No | | 0.10 | No | Yes | No | 0.01 |
| No | No | Yes | | 0.00 | No | No | Yes | 0.00 |
| No | No | No | | 0.30 | No | No | No | 0.16 |
|  | | | | | | | | | | | | | |
| **p(Tx, DNR, S)** | | | | **Eliminate Severity, S** | **p(S, Tx)** | | | | **Add to One** | **p(S, Tx)** | | |  |
| **DNR** | **S** | **Tx** | **p** | **S** | **Tx** | | **p'** | **DNR** | **Tx** | **p** |
| Yes | Yes | Yes | 0.03 | Yes | Yes | | 0.05 | Yes | Yes | 0.18 |
| Yes | Yes | No | 0.06 | Yes | No | | 0.07 | Yes | No | 0.25 |
| Yes | No | Yes | 0.00 | No | Yes | | 0.00 | No | Yes | 0.00 |
| Yes | No | No | 0.00 | No | No | | 0.16 | No | No | 0.57 |
| No | Yes | Yes | 0.02 |  | | | | | | | | |
| No | Yes | No | 0.01 |
| No | No | Yes | 0.00 |
| No | No | No | 0.16 |
|  | | | | | | | | | | | | | |
| **p(S, Tx)** | | | **Joins with** | **p(O|Tx, S)** | | | | | **Equals =** | **p(O, Tx, S)** | | | |
| **DNR** | **Tx** | **p** | **Tx** | **S** | **O** | | **p** | **Tx** | **S** | **O** | **p** |
| Yes | Yes | 0.18 | Yes | Yes | + | | 0.20 | Yes | Yes | + | 0.18 |
| Yes | No | 0.25 | Yes | Yes | - | | 0.10 | Yes | Yes | - | 0.09 |
| No | Yes | 0.00 | Yes | No | + | | 0.10 | Yes | No | + | 0.13 |
| No | No | 0.57 | Yes | No | - | | 0.00 | Yes | No | - | 0.00 |
|  | | | No | Yes | + | | 0.10 | No | Yes | + | 0.00 |
| No | Yes | - | | 0.30 | No | Yes | - | 0.00 |
| No | No | + | | 0.10 | No | No | + | 0.30 |
| No | No | - | | 0.10 | No | No | - | 0.30 |

*Note*: Tx = Treatment, P = probability, S = severity of illness, DNR = do not resuscitate, O = outcome.

Let us assume that we want to estimate the effect of treatment on outcome for severely ill patients. In the first step, we join the severity and the conditional DNR tables to estimate the joint probability of DNR and severity of illness. The tables are joined on variables they share (in this case the severity of the illness). The values of the new joint table are provided by multiplying each conditional probability by the prior probability of the condition and making sure that the conditional probabilities add up to 1. For example, the first row of the new joint table is provided by multiplying the probability of being severely ill (0.4) by the conditional probability of signing a DNR order among severely ill patients (0.8). No elimination is necessary at this point, as both DNR and severity of illness are used in selection of treatment. We proceed to join the joint probability of DNR and severity of illness with the conditional probabilities of treatment.

Because we want to examine the impact of treatment on outcome for severe patients, we are not interested in the DNR orders. We can eliminate this variable by summing across it. In exhibit 20.8, we see this done in the third row of the table. The DNR variable is dropped and the cell values for the same combination are added to each other. If through elimination, the probabilities do not add up to 1, then each cell value is divided by the total so that probabilities are forced to add up to one.

If a variable is substantiated—meaning that it has occurred—we remove all rows corresponding to unsubstantiated levels of the variable. Then we only join all substantiated variables.

# [H1] Causal Impact

The previous steps demonstrated how to predict the probability of an event in the network. The probability of an event is not the same as the causal impact of the network on the event. Probability is an association measure. To have causal interpretation, we need concepts that go beyond the probabilities of events. When we talk about “cause and effect,” we refer to situations in which, if we take an action, we will see the effect. If we do not, we will not. A causal impact is verified by manipulating the world—making a change and examining the outcomes after and before the change. Likewise, in a causal network, the causal impact is revealed by surgically changing the network, d-separating the relationship we are interested from the rest of the network. Pearl called this network manipulation the “do operation” to emphasize that we want to see the impact of actions and not the probabilities of events.

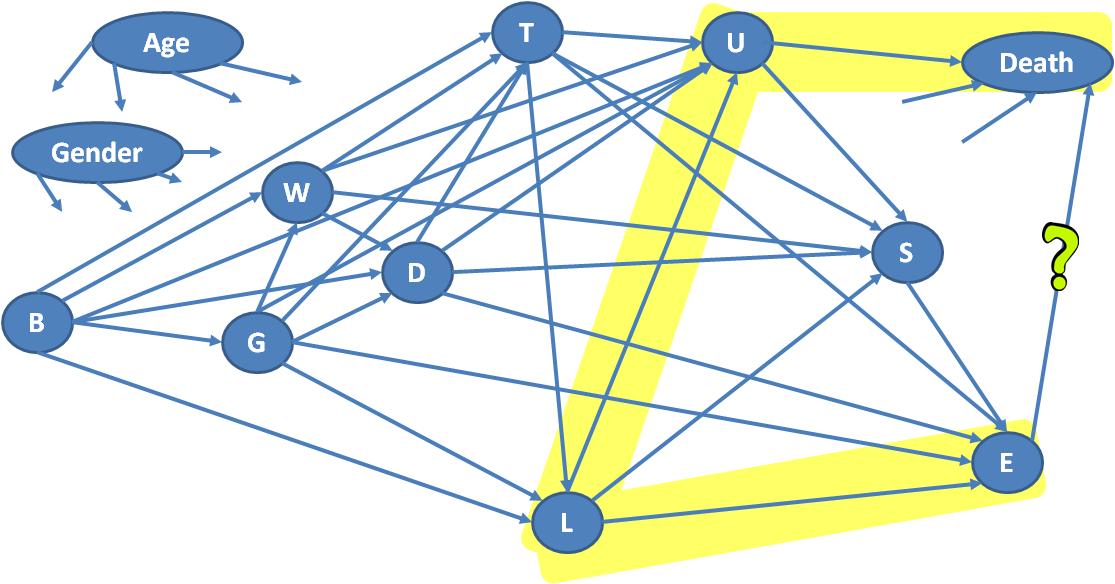
For ease of communication, we will talk about two variables in the network: treatment and outcome. The do operation calculates the causal impact of treatment on outcome. We start to directionally separate the relationship between treatment and outcome from the rest of the network. Once separated, the difference between treated and untreated group is no longer confounded with other variables. This difference provides the causal impact of treatment on outcome.

## [H1] Back-Door Paths and Markov Blankets

To analyze the causal relationship between treatment and outcome, analysts often refer to a *back-door path*. A back-door path starts from the outcome, reaching back—and pointing—to treatment. It is a series of nonintersecting adjacent edges that start from the outcome and go back to, and cause, the treatment. In establishing the path, direction of the arcs does not matter, except at the very last arc, which points to treatment. Two variables are independent if there are no paths between them or if paths between them are blocked through stratification. Every variable on the back-door path is called a *covariate* because it is associated directly, or indirectly, with both treatment and outcome. All covariates of treatment and outcome are on one or more back-door paths.

In exhibit 20.9, we show a network of disabilities and the back-door path from death (the outcome) to eating disability (the treatment variable). In this network, the effect of age and gender are not shown completely to make the display easier to understand. Furthermore, not all relationships to survival are shown, again to reduce the number of arcs displayed. Death is the outcome. Suppose we want to examine the effect of inability to eat on death. In this context, inability to eat is a treatment/exposure variable and we want to understand if people who are unable to eat are more likely to die. We show a question mark on the link between inability to eat and death as we want to estimate this impact. In exhibit 20.9 the back-door path starts with death and moves against the arc direction to urine incontinence. Then, moving again against the direction of the arc, it reaches bowel incontinence, and finally goes in the direction of the arc to eating disabilities. All back-door paths must include a parent, and not a child, of treatment/exposure. All nodes on the path—in this case urine and bowel incontinence—are covariates of treatment/exposure and outcome. The mere fact that urine and bowel incontinence are on the back-door path is sufficient for us to deduce that they are also correlated (directly or indirectly) with eating disorder and death.

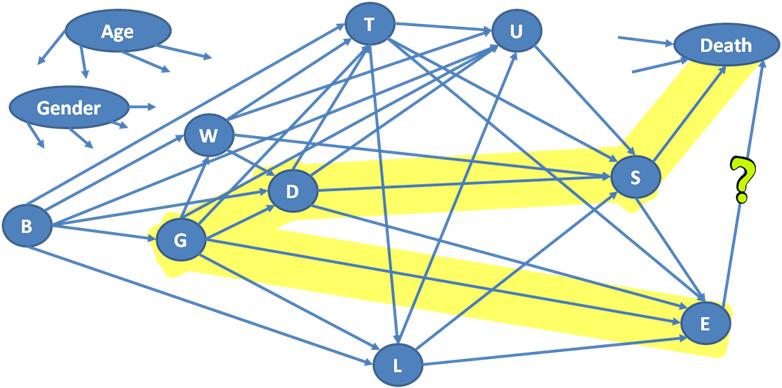
**Exhibit 20.9** Example of a Back-Door Path for the Impact of Eating Disability on Death



*Note*: For ease of understanding, not all directed arcs are shown. B = unable tobathe, W = unable to walk, G = unable to groom, D = unable to dress, T = unable to toilet, L = bowel incontinent, S = unable to transfer, U = urinary incontinent, E = unable to eat.

In exhibit 20.10, we show a different path starting from death, the outcome. It also points to eating disorder, the treatment/exposure variable. Yet, it goes through an entirely different set of covariates: unable to transfer, unable to dress, and unable to groom. These three variables are also covariates that confound the impact of the treatment/exposure variable on the outcome. Any variable on any back-door path from outcome to treatment/exposure is associated with both the outcome and treatment/exposure and therefore is a covariate. All covariates should be controlled before the effect of treatment/exposure can be accurately estimated.

**Exhibit 20.10** Another Example of Back-Door Path from Death to Eating Disability



*Note*: For ease of understanding, not all directed arcs are shown. B = unable tobathe, W = unable to walk, G = unable to groom, D = unable to dress, T = unable to toilet, L = bowel incontinent, S = unable to transfer, U = urinary incontinent, E = unable to eat.

Back-door paths can be discovered from the correlation matrix and knowledge of the sequence of variables. In the correlation matrix, the path includes pairs of overlapping correlated variables that start with outcome and reach to a variable that precedes treatment and is correlated with treatment. Suppose that all covariates occur prior to treatment and treatment prior to outcome. Then the covariates that have large and statistically significant correlation with treatment (but not with each other) are initial candidates for parents in the Markov Blanket. If relying on the correlation matrix and sequence of variables, the determination of a back-door path does not require knowledge of the network structure. Therefore, back-door paths can be identified in settings where no network model can be constructed.

**[H2] Blocking Back-door Paths**

Identifying the back-door paths is helpful in understanding the covariates that should be controlled for before an unconfounded impact of treatment/exposure on outcome can be assessed. If all back-door paths are blocked, then the relationship between treatment/exposure and outcome is said to be directionally separated (d-separated) from the rest of the network; meaning that we can study the impact of treatment/exposure on outcome without paying attention to the rest of the network.

The concept of blocking back doors was first used in causal network models (Pearl 2000). *Blocking the back-door paths* is stratifying at least one variable on the back-door path, so that the covariates on the path cannot affect both treatment and outcome. They can affect one or the other but not both treatment and outcome. The stratification breaks the correlation between covariates and either treatment/exposure or outcome. For the paths in exhibit 20.10, stratifying S will break the path; stratifying L will break the path in exhibit 20.9 but not in exhibit 20.10. Different nodes break different paths. Since the network has many back-door paths, a strategy is needed to block all paths with as few stratification as possible.

The problem of identifying an optimal set of variables to break back-door paths is further complicated by the fact that sometimes blocking one path may open new ones. Stratifying a variable may lead to new paths when the variable is a common effect of two or more causes. The stratification introduces new correlations between the causal variables. These new pathways may reestablish a covariate or make other variables into new covariates. Therefore, as the analyst blocks a back-door path, she must recheck that no new pathways have been introduced.

**[H2] Markov Blankets**

The Markov blanket is a minimum set of variables that would block all back-door paths. The Markov blanket of a variable includes its parents, children, and co-parents. For example, the Markov blanket of eating disorder in Exhibit 20.10 includes all of its parents (S, T, D, G, and L), all of its children (death) and all of its co-parents (parents of death, which include U and other variables not shown). Parents in the Markov blanket separate the impact of treatment/exposure on outcome from the rest of the network. Therefore, this concept can be used to select a smaller set of relevant features in high-dimensional problems. The Markov blanket has proven very effective for feature reduction in high-dimensional problems, sometimes reducing the number of variables a thousand fold without any loss of accuracy (Aliferis, Tsamardinos, and Statnikov 2003; Fu and Desmarais 2010; Shen, Li, and Wong 2008; Tan and Zhifa 2013; Zeng, Jian, and Lin. 2009).

The establishment of the Markov Blanket is also very useful in creating the network. Parents in the Markov blanket are the arcs that are drawn in a causal network—therefore, establishing parents in a Markov blanket also specifies the structure of the entire network. If one identifies the Markov Blanket of all variables, then one can easily draw the network. Likewise, if one has a drawing of the network, then the Markov Blanket of each variable can be read from the drawing.

# [H1] Estimating Structure and Parameters of Causal Networks

To specify a causal network, one has to identify three things: (1) the pairs of variables that directly affect each other, (2) the direction of the causality, and (3) the magnitude of the impact of the causes. The first two steps are referred to as the *identification of the network structure* and the last one as the *specification of network parameters*.

## [H1] Learning Associations among Pairs of Variables

There are three types of algorithms for learning the association among variables: search and score, constraint based, and multivariate. This chapter briefly describes each algorithm and then focus on the use of regression in learning the structure of causal networks.

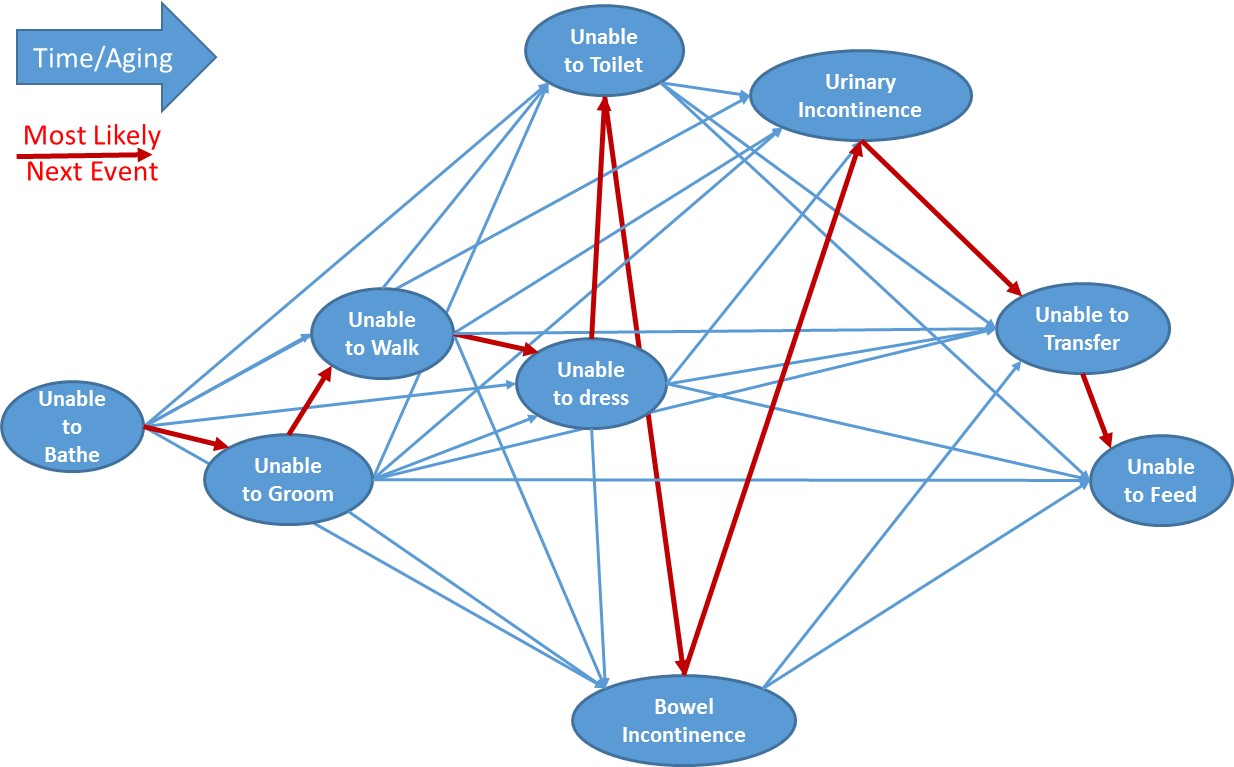
### [H2] Search-and-Score Methods

The first set of algorithms includes taboo, maximum-minimum hill climbing, and restricted maximization algorithms (e.g. Friedman, Nachman, and Peér 1999; Heckerman, Geiger, and Chickering 1995; Jouffe and Munteanu 2001; LaMunteanu and Bendou 2001; Naïm et al. 2011)[[1]](#endnote-1),[[2]](#endnote-2),[[3]](#endnote-3),[[4]](#endnote-4),[[5]](#endnote-5),[[6]](#endnote-6),[[7]](#endnote-7),[[8]](#endnote-8),[[9]](#endnote-9),[[10]](#endnote-10),[[11]](#endnote-11),. In these algorithms, various possible network structures are scored, and the structure that best fits the data is chosen. As you can imagine, when the number of variables in a network is large, the possible interactions among the variables get very large, and search-and-score algorithms do not work efficiently.

A recent study can show how search-and-score methods can be used to create a network model (Levy et al. 2016). The data included disabilities of 296,051 residents nursing homes run by the US Department of Veterans Affairs (VA). The researchers examined veterans’ experiences from January 1, 2000, through September 10, 2012. They learned the direction of cause and effect in the network using search-and-score techniques, facilitated with information on sequence among the disabilities.

Several software are available to learn network models from data. We used BayesiaLab (version 5.3). We constructed a network using five different learning algorithms: max spanning tree, Taboo, EQ, SopLeq, and Taboo Order, all subject to sequence constraints. Among the five learned networks, we chose the one with best fit to 40 percent of data randomly set aside for validation. Exhibit 20.11 shows the resulting network for the nine disabilities. Causal networks do not allow circular networks; therefore, we did not include recovery from a disability in this analysis. The sequence of occurrence of disabilities goes from left to right, with events occurring more to the left being more likely to occur first. The most likely path for disabilities is shown as a dashed line. Keep in mind that this directed network structure is not causal, in the sense that we do not think that one disability causes another. Future disabilities are likely to occur because of diseases and aging and not previous disabilities. Yet the model is helpful in understanding how directed acyclical graphs can be used to describe networks of events. In this network, disabilities not directly linked to each other are still associated with each other but transition between them is mediated by other disabilities. The red line shows the most common path for transitions among disabilities.

**Exhibit 20.11** A Network Model of Progression of Disabilities

****

**[H2] Constraint-Based Algorithms**

The second set of algorithms, the constraint based algorithms, relies on tests of conditional independence (Aliferis, Tsamardinos, and Statnikov 2003; Cheng et al. 2002; Margaritis 2003; Pearl 1988; Spirtes, Glymour, and Scheines 2000; Verma and Pearl 1990). There are many network structures that can be detected from a particular pattern of independence. For example, a causal chain A🡪 B 🡪 C can be detected by the following set of conditional independence tests:

* A and B are dependent.
* B and C are dependent.
* A and C are likely to be dependent; conditional on B, A and C are independent.

Similarly, one can recognize a common effect (i.e., A 🡪 B 🡨 C) from the following set of conditional independence tests:

* A and B are dependent.
* C and B are dependent.
* A and C are independent; conditional B, they are dependent.

A good example of constraint-based algorithms is the grow-shrink algorithm. This algorithm has two phases. In the grow phase, the strongest relationships are used to connect pairs of variables. In the shrink phase, the relationships between selected pairs of variables are reexamined, and pairs that no longer have a strong relationship are dropped from the analysis.

### [H2] Multivariate Methods

Finally, the multivariate methods for discovery of network structure rely on ordinary, logistic, LASSO or Poisson regression or on correlations (Agresti 2002; Allen and Liu 2013; Aragam and Zhou 2015; Fu and Zhou 2013; Han and Zhong 2016; Park and Raskutti 2015). The multivariate methods are particularly advantageous because statisticians are familiar with them. It is a new use for a familiar regression tool. Some of these methods (e.g., correlation or Poisson regression) are efficient and can easily be used in massive, high-dimensional data.

Shojaie and Michaildis used LASSO regression to learn the network structure (Shojaie and Michaildis 2009). *Lasso regression* is a type of regression that requires the effect size to be larger than a cutoff value. These methods of regression are especially effective in large data sets such as data from EHRs. In massive data, everything is statistically significant, and methods are needed that would focus on large effect sizes and ignore small effect size. In these regressions, the response, or dependent, variable is any variable in the network. The independent variables are all the variables that precede the response variable. If nothing occurs prior to their response variable, some regression will have no independent variables. Other regressions may have many independent variables when many variables occur prior to the response variable. Here are the steps that should be followed to identify the parents in the Markov Blanket of a variable, say the treatment/exposure variable:

1. First, remove from analysis all variables that occur after treatment/exposure. This is done because the impact of treatment on outcome is distorted, if we stratify variables on the causal path of treatment to outcome. Furthermore, we are only interested in detecting parents in the Markov Blanket so all children, and by extension co-parents, can be ignored.
2. Second, the treatment/exposure variable is regressed on main effects of all covariates that occur prior to it. The variables that have a statistically significant relationship to treatment/exposure variable, referred to as Zi, are the initial set of candidates for the parents in Markov blanket of treatment/exposure.
3. Third, the regression is expanded to include interaction of variables outside of the blanket with the Zi variables. The idea is to verify that no variables outside of the blanket would affect treatment/exposure if the parents were to be stratified. Interaction effects show the effect of outside variable when Zi = 1.

A summary of these steps are provided in exhibit 20.12. This exhibit shows how Pearl defined blocking back-doors and how Shojaie’s method was adjusted to accomplish Pearl’s steps.

**Exhibit 20.12** Using LASSO Regression to Identify Parents in Markov Blanket

|  |  |
| --- | --- |
| **Pearl’s Steps** | **Steps in Using Regression** |
| 1. Remove all descendants of X, except Y. Stratifying the variables on causal path from X to Y distorts the impact of X on Y. | 1. Remove all variables that occur after X. These variables cannot be parents of X |
| 1. Block all spurious back-door paths from X to Y but leave all directed paths unperturbed. In a causal network, back-door paths are collection of pairs of overlapping nodes that are associated with each other (have an arc between them) and end up with a direct arc to X. | 1. Regress Y on all covariates that occur before X. The variables Zi that have a large and statistically significant impact on X are parents in the Markov blanket of X. |
| 1. Verify that no new spurious paths have been identified. Stratifying/blocking a common effect will open new spurious correlations in the data that could affect the treatment. | 1. Add to parents of X, any variable that interacts with Zi and has a large and statistically significant impact on X. This is a new relationship that occurs only when Zi is stratified and is equal to 1. |
| 1. Condition on nodes that block all back-door paths. One option is to condition on all nodes that are parents of X. | 1. Stratify all parents of X. |

## [H1] Directing the Arcs in the Network

All network learning algorithms, with some exceptions (Aliferis, Tsamardinos, and Statnikov 2003), determine the direction of the arcs after establishing the network structure. Judea Pearl suggests that some arcs can be directed using conditional independence tests and remaining arcs can be oriented randomly as long as no cycles are created. The test of cycles requires knowledge of network structure; hence, many algorithms orient arcs after establishing the structure. The approach to orientation of arcs is surprisingly flippant given the effort to identify causes. At times, if the order cannot be determined, the advice is to choose an order randomly that does not cause cycles. I prefer to observe the time of occurrence of the variables, a piece of information readily available in EHRs. Every event in an EHR comes with a date and time stamp, so it is easy to establish the sequence of events for a patient.

There are many methods for establishing sequenceamong pairs of variables.

1. Pearl’s collider test is one method. These tests establish a common effect through conditional independence tests. The existence of common effect directs the causes towards the effect. All other variables are randomly assigned so that no new cycles occur in the network.
2. *Use the definition of the variables.* Many statisticians already do so. They design studies, so that some variables are measuring events at baseline, other variables refer to treatment after baseline, and outcomes are measured last. These assumptions about timing of events allow an easy way to establish a partial sequence. For example, by definition, race occurs at birth. It is established prior to medical history, which occurs before current conditions, which occurs prior to treatment, which also occurs prior to outcomes. Selecting the variables carefully would, by design, create a partial sequence among the variables.
3. *Use error reduction*. One could employ Goodman and Kruskal (1954) error reduction for predicting A from B or B from A. Most statistical measures are symmetric measures and cannot be used to order a pair of variables. This is not the case for Goodman and Kruskal error reduction. Vang (2008) used this method to examine sequence among predictors of drug abuse.
4. *Use strength of causal reasoning*. Use a model of how humans think through causality. For example, Zargoush and colleagues (2014) used probabilistic contrast model, a model of how humans judge strength of causality, to sequence the variables.
5. *Use the age at which the variable typically occurs*. In this approach, events that occur at a later age are considered to occur after events that occur ata younger age. This method can be used in cross-sectional data. For example, a number of cross-sectional studies show that feeding disabilities occur after walking disabilities because they occur at a later age (Levy et al. 2016)
6. *Use longitudinal order*. In this approach, the data and time of occurrence of the events is used to identify if one event occurs prior to another. This approach is the gold standard for defining sequence among variables. These data are widely available in EHRs, where events are time-stamped. Even when relationships among variables are examined in cross-sectional data, the longitudinal order of the variables can be separately measured and used to inform the algorithm.

Sequence information is widely available. Multiple methods exist to extract it. Analysts who construct causal networks focus mostly on the common effect test, which requires knowledge of the network structure. Obviously, this test cannot be used to improve the learning of network structure. Other methods of learning sequence do not require knowledge of the network structure and therefore can be implemented before the structure is discovered.

A recent study by Alemi and colleagues has shown that if the sequence among the variables is well established, then nearly all algorithms, whether constraint based, search and score, or multivariate, become more accurate. In this study, performance of eight network learning algorithms were examined with or without use of sequence constraints. For all eight algorithms, the Area under the Receiver Operating Curve (AROC) of the sequenced algorithm was significantly higher than the AROC of the unsequenced algorithms. Furthermore, when sequence was used all algorithms had near perfect accuracy, so it seemed that the algorithms did not matter. High accuracy rates in learning the network structure was achieved no matter what algorithm was used.. These data highlight the advantage of establishing sequence through methods other than common effect technique and of using the resulting sequence to improve the performance of network learning algorithms.

Often, the sequence of occurrence between a pair of variables is not deterministic. Sequence is an uncertain judgment that must arise from data. Naturally, these inferences could be erroneous. Also, Sequences derived from data will almost never apply to all individuals. The majority of patients may experience one sequence of variables, while a minority may experience the reverse order. For example, consider substance abuse disorders, heart attack, and Alzheimer’s disease. Most patients will report substance abuse at younger age than they experience a heart attack. In addition, the majority of patients will report Alzheimer’s disease at a later age than heart attack. Despite these patterns, it is also possible that some patient have these diseases in reverse order, meaning that substance abuse can occur after heart attack or Alzheimer’s disease could occur before heart attack.

## [H1] Learning the Parameters of the Network

Once the structure of the network is understood, the joint distribution of the variables in the network can be easily established. In causal networks, the joint probability of all events can be read directly from the structure of the graph. For each variable, one identifies the parents of the node from the graph. Then the joint probability of events in the network is calculated by conditioning on the parents of the variable, in a formula that looks like

In this equation, are *n* variables inside the network; is the joint probability of observing the combination of the varaibles ; indicates the parents of the variable. The sign indicates the product of values calculated for each variable, going from the first to the nth variable. This equation radically simplifies the estimationof the distribution of the variables and distribution of the combination of variables. It says that only parents matter, everything else in the network can be ignored.

## [H1] Verification

Pearl showed that stratifying variables can open new back-door paths between outcome and treatment. Therefore, it is important to verify that stratifying parents to treatment does not open new pathways. To verify, we need to stratify all candidate variables and examine the impact of the variables that were not stratified one at a time. If a variable that previously had no impact now has an impact, a new back-door path has been created and the variable must also be stratified. A quick way to accomplish this task is to use stratified covariate balancing. For example, to see whether there are new back-door paths to treatment (T) after stratifying the parents to treatment into *k* strata, we will calculate the common odds ratio, for impact of an unstratified covariate, U, on treatment as

,

where the constants *a*s through *d*s are defined in each stratum *s* as

* *as* is the count of times when both U and T are present in stratum *s*
* *bs* is the count of times when U is present and T is absent in stratum *s*
* *cs* is the count of times when U is absent and T is present in stratum *s*
* *ds* is the count of times when both U and T are absent in stratum *s*
* and

## [H1] Calculating Causal Impact of Cancer on Survival

To demonstrate the calculation of causal impact, we examined the survival time of veterans in 168 medical centers over seven years starting in January 2008 and ending in December 2015. To be included in the cohort, patients had to have two primary care visits, not more than two years apart, to a VA facility. Of the pool of subjects, 4,681,809 had at least two primary care visits. Among these, we focused on survival rates for 829,827 hospitalized veterans. The independent variables were comorbidities of stomach cancer. There were 10,292 unique comorbidities among the patients. We excluded any diagnoses that occurred after stomach cancer, as these were considered possible complications of cancer and potentially on the causal path from cancer to survival. We also excluded any diagnoses that did not occur at least 100 times in patients who had stomach cancer. Examples of common diagnoses included “history of smoking” and “concurrent malignant neoplasm of the esophagus.” A complete list is provided in exhibit 20.13.

**Exhibit 20.13** Comorbidities of Stomach Cancer

|  |  |  |
| --- | --- | --- |
| **Code** | **Diagnoses** | **Frequency** |
| 403.90 | Hypertensive renal disease | 145 |
| 427.31 | Atrial fibrillation | 244 |
| 428.0 | Congestive heart failure | 156 |
| 600.00 | Benign hypertrophy of prostate without urinary obstruction | 224 |
| 585.9 | Chronic kidney disease | 131 |
| 414.00 | Coronary atherosclerosis | 101 |
| 599.0 | Other disorders of urinary tract | 158 |
| 414.01 | Coronary atherosclerosis | 279 |
| 244.9 | Hypothyroidism | 142 |
| V66.7 | Encounter for palliative care | 151 |
| 584.9 | Acute kidney failure | 231 |
| 578.9 | Hemorrhage of gastrointestinal tract | 117 |
| 197.7 | Malignant neoplasm of liver | 197 |
| V58.61 | Long-term (current) use of anticoagulants | 111 |
| 486. | Pneumonia | 146 |
| 458.9 | Hypotension | 125 |
| 496. | Chronic airway obstruction | 344 |
| 285.9 | Anemia | 300 |
| 280.9 | Iron-deficiency anemia | 195 |
| 272.4 | Other and unspecified hyperlipidemia | 572 |
| 564.00 | Constipation | 159 |
| 787.20 | Dysphagia | 189 |
| 401.9 | Unspecified essential hypertension | 928 |
| 511.9 | Unspecified pleural effusion | 112 |
| V15.82 | Personal history of nicotine dependence | 129 |
| 276.51 | Dehydration | 162 |
| 263.9 | Unspecified protein-calorie malnutrition | 134 |
| 530.81 | Esophageal reflux | 478 |
| 276.8 | Hypopotassemia | 148 |
| 276.1 | Hypo-osmolality and hyponatremia | 114 |
| 150.9 | Malignant neoplasm of esophagus | 123 |
| E849.7 | Unspecified place in other specified residential institution | 142 |
| 311. | Depressive disorder | 199 |
| 309.81 | Post-traumatic stress disorder | 138 |
| 305.1 | Tobacco use disorder | 356 |

*Note*: Diagnoses listed are from International Classification of Diseases, version 9.

The main-effect LASSO regression of treatment on covariates that precede treatment is provided in exhibit 20.13. We stratified variables that were on the parents in the Markov blanket of treatment. Eleven comorbidities had a statistically significant relationship with stomach cancer and had an effect size that exceeded odds of 1.5. We stratified these eleven variables. We also verified whether any other variables besides these eleven had a statistically significant relationship with stomach cancer once we stratified them. No new relationships were observed, and therefore these eleven variables are blocking all back-door pathways to stomach cancer. Merely stratifying these eleven variables is sufficient to accurately measure the causal impact of stomach cancer on survival.

**Exhibit 20.13** Identifying Parents in Markov blanket through Lasso Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Inpatient Diagnoses Codes** | **Regression of Stomach Cancer on Prior or Current Comorbidities** | | |
| **Effect Size** | **P-Value** | **Parents to Treatment** |
| 309.81 | -0.861 | 0 | Yes |
| 150.9 | 2.813 | 0 | Yes |
| 263.9 | 1.595 | 0 | Yes |
| 276.51 | 0.736 | 0 | Yes |
| 511.9 | 1.443 | 0 | Yes |
| 787.20 | 1.549 | 0 | Yes |
| 280.9 | 1.669 | 0 | Yes |
| 197.7 | 2.966 | 0 | Yes |
| 578.9 | 1.339 | 0 | Yes |
| V66.7 | 1.496 | 0 | Yes |
| 585.9 | -0.91 | 0.038 | Yes |
| 403.90 | 0.701 | 0.01 | Yes |
| 305.1 | -0.28 | 0.027 | No |
| E849.7 | 0.546 | 0.004 | No |
| 276.1 | 0.457 | 0.048 | No |
| 530.81 | 0.636 | 0 | No |
| V15.82 | 0.509 | 0.013 | No |
| 401.9 | 0.194 | 0.035 | No |
| 564.00 | 0.627 | 0.001 | No |
| 272.4 | -0.287 | 0.028 | No |
| 285.9 | 0.641 | 0 | No |
| 427.31 | 0.485 | 0.005 | No |

# *Note*: Effect size should be higher than +.4 or lower than -.4 to be eligible as a parent.

# [H1] Calculating the Causal Impact of Eating Disability on Death

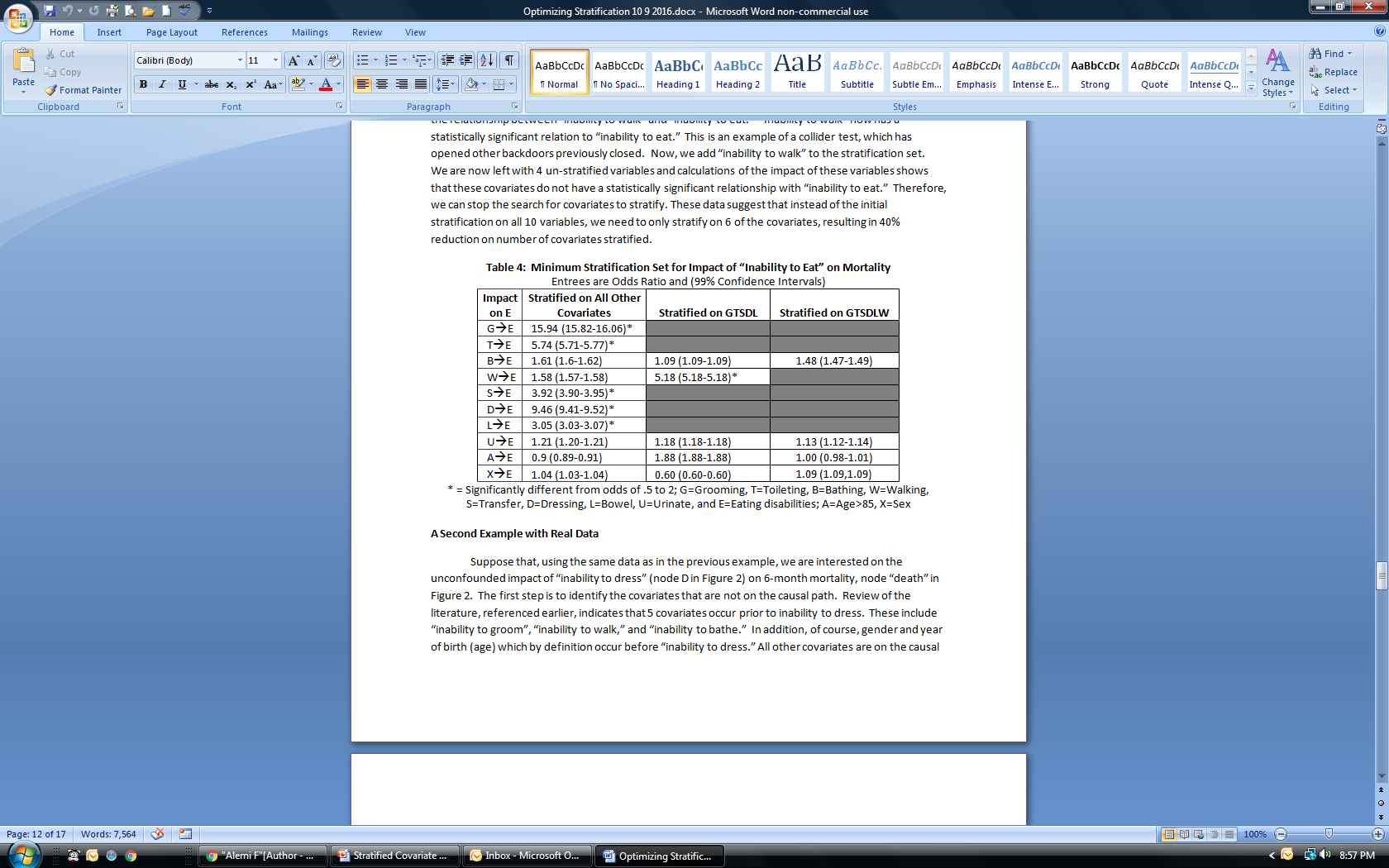
For a second example, we look at how back-door paths were blocked for evaluating the causal impact on six-month mortality of the onset of eating disability. The sample included 296,051 residents in VA nursing homes called community living centers. The study ran from January 1, 2000, through September 10, 2012. These data include a comprehensive assessment of residents in the following domains:

* Cognition
* Communication and hearing
* Vision
* Physical functioning
* Continence
* Psychosocial well-being
* Mood and behavior
* Activity pursuit patterns
* Disease diagnoses
* Other health conditions (i.e., accidents, pain)
* Oral and nutritional status
* Oral and dental status
* Skin condition
* Medication use
* Treatments and procedures
* Patterns of activity

By policy, assessments are to be done within 14 days of admission, at least quarterly—or sooner when there has been an event such as hospitalization or when the nursing home staff identifies a change in the resident’s status. In our data, there were two peaks in the distribution of assessments—one for residents assessed every month (75,994 residents) and the other for residents assessed every three months (42,904 residents). The average time between assessments was 115 days, and the standard deviation was 235 days. Data were used to classify the patient as having eating (E), bathing (B), grooming (G), dressing (D), toilet use (T), transfers (S), and walking (W) disabilities and bowel (L) and urine incontinence (U). We were interested in verifying the causal impact of eating disability on mortality.

Exhibit 20.14 shows the steps undertaken to block every parent in the Markov blanket of eating disability, shown as E. When all variables were stratified, grooming, toileting, transferring, and dressing disabilities, as well as bowel incontinence, had a statistically significant and large effect size (i.e., effect size > 1.5 or effect size < 1 ÷ 1.5)); see Exhibit 20.17 for details. Therefore, we concluded that these five variables were parents on the Markov blanket of eating disability. We then stratified these five variables and reconsidered whether any other variables were now related to eating disability (i.e., had both a statistically significant effect and a large effect). Surprisingly, a back-door had opened with walking. Walking was added to the list of variables to be stratified. We reexamined the relationships between remaining variables and eating disorders while stratifying six variables, and no new relationships were found.

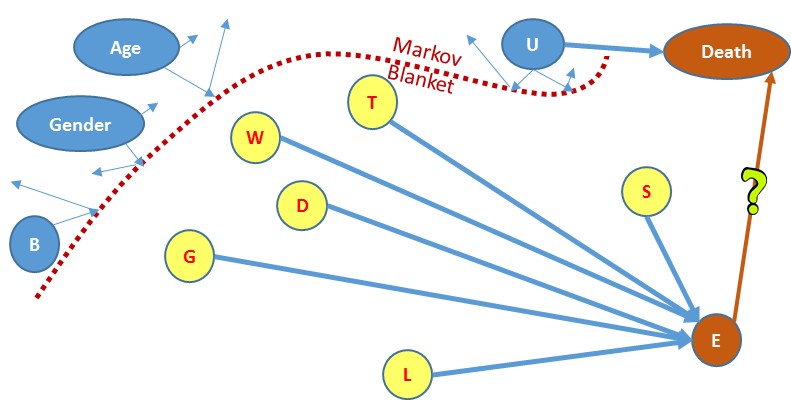
**Exhibit 20.14** Blocking Back-Doors from Mortality to Eating Disorders



*Note*:E = eating, B = bathing, G = grooming, D = dressing, T = toileting, S = transfer, W = walking disabilities, L = bowel incontinence, U = urine incontinence, A = old age, X = male gender.

The dashed line in Exhibit 20.15 shows the parents in a Markov blanket of eating disability (node E). In this analysis, eating disorder is an exposure/treatment variable and mortality is an outcome. Six variables (grooming, dressing, toileting, transfer, walking disabilities, and bowel and urine incontinence) were parents to eating disability and can be stratified. If stratified, they break the link between all variables and eating disability. In particular, these six variables blocked the effects of bathing disability, urine incontinence, age, and gender. When these links are broken, then these variables are no longer a covariate as none affect both eating disorders and mortality. We show these blocked relationships as arcs reverberating from the blanket so we could emphasize that these relationships no longer exist. Note that exhibit 20.15 does not show the relationship among the six parents inside the blanket, as these variables are assumed to be substantiated and stratified; they are either present or absent; and are not predicted from each other. The six variables are related to each other but these relationships are immaterial now that they are stratified. Also note that urine incontinence is a co-parent to eating disorder and inside the Markov blanket but it is not among the list of variables that should be stratified, only parents in the Markov blanket should be stratified.

**Exhibit 20.15** Parents in Markov Blanket of Eating Disabilities

****

*Note*:Not all arcs in the network are shown to simplify the display. E = eating, B = bathing, G = grooming, D = dressing, T = toileting, S = transfer, W = walking disabilities, L = bowel incontinence, U = urine incontinence.

# [H1] Summary

This chapter introduced causal network, including the concept of directional separation, Markov blankets, and back-door paths. We showed how a network can calculate the probability of an event. We also showed how the network could identify the causal impact of a variable. Although this chapter introduced many novel ideas and new terminology, the methods of analysis relied on the familiar tools of regression and stratification.

# [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are on the course website.

# [H1] References

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ: Wiley.

Alemi F, Avramovic S, Schwartz MD. Electronic Health Record-Based Screening for Substance Abuse. Big Data. 2018 Sep 1;6(3):214-224.

Aliferis, C. F., I. Tsamardinos, and A. Statnikov. 2003. “HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection.” *AMIA Annual Symposium Proceedings* 2003: 21–25.

Allen, G. I., and Z. Liu. 2013. “A Local Poisson Graphical Model for Inferring Networks from Sequencing Data.” *IEEE Transactions on Nanobioscience* 12 (3): 189–98.

Aragam, B., and Q. Zhou. 2015. “Concave Penalized Estimation of Sparse Gaussian Bayesian Networks.” *Journal of Machine Learning Research* 16: 2273–328.

Atwood A, Lo Sasso AT. The effect of narrow provider networks on health care use. J Health Econ. 2016 Dec; 50: 86-98.

Bang S, Kim JH, Shin H. Causality modeling for directed disease network. Bioinformatics. 2016 Sep 1; 32(17): i437-i444.

Cheng J., R. Greiner, J. Kelly, D. A. Bell, and E. Liu. 2002. “Learning Bayesian Networks from Data: An Information-Theory Based Approach.” *Artificial Intelligence Journal* 137: 43–90.

Friedman, N., E. Nachman, and D. Peér. 1999. “Learning Bayesian Network Structure from Massive Datasets: The “Sparse Candidate” Algorithm.” In *Proceedings of the* *Fifteenth Conference on Uncertainty in Artificial Intelligence*, edited by K. Laskey and H. Prade, 206­–15. Los Angeles: Morgan Kaufmann.

Fu, S., and M. C. Desmarais. 2010. “Markov Blanket Based Feature Selection: A Review of Past Decade.” In *Proceedings of the World Congress on Engineering*, edited by S. I. Ao, L. Gelman, D. W. L. Hukins, A. Hunter, and A. M. Korsunsky, 321–28. Hong Kong, China: Newswood Ltd.

Fu, F., and Q. Zhou. 2013. “Learning Sparse Causal Gaussian Networks with Experimental Intervention: Regularization and Coordinate Descent.” *Journal of the American Statistical Association* 108 (501): 288–300.

Goodman, L. A., and W. H. Kruskal. 1954. “Measures of Association for Cross Classifications.” *Journal of the American Statistical Association* 49: 732–64.

Han, S. W., and H. Zhong. 2016. “Estimation of Sparse Directed Acyclical Graphs for Multivariate Counts.” *Biometrics* 72 (3): 791–803.

Heckerman, D., D. Geiger, and D. Chickering. 1995. “Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.” *Machine Learning* 20 (3): 197–234.

Jouffe, L., and P. Munteanu. 2001. “New Search Strategies for Learning Bayesian Networks.” In Proceedings of Tenth International Symposium on Applied Stochastic Models and Data Analysis, edited by G. Govaert, J. Janssen, and N. Limnios, n. p. Compiègne, France: UTC.

Katz, S., A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe. 1963. “Studies of Illness in the Aged: The Index of ADL: A Standardized Measure of Biological and Psychological Function.” *JAMA: Journal of the American Medical Association* 185 (12): 914–19.

Kheirbek RE, Beygi S, Zargoush M, Alemi F, Smith AW, Fletcher RD, Seton PN, Hawkins BA. Causal Analysis of Emergency Department Delays. Qual Manag Health Care. 2015 Jul-Sep;24(3):162-6.

Levy, C. R., M. Zargoush, A. E. Williams, A. R. Williams, P. Giang, J. Wojtusiak, R. E. Kheirbek, and F. Alemi. 2016. “Sequence of Functional Loss and Recovery in Nursing Homes.” *Gerontologist* 56 (1): 52–61.

Margaritis, D. 2003. “Learning Bayesian Network Model Structure from Data.” Ph.D. diss., Carnegie Mellon University.

Munteanu, P., and M. Bendou. 2001. “The EQ Framework for Learning Equivalence Classes of Bayesian Networks.” In *Proceedings of the 2001 IEEE International Conference on Data Mining,* 417– 424.

Naïm P, P. H. Wuillemin, P. Leray, O. Pourret, and A. Becker. 2011. *Réseaux Bayésiens*. Paris: Eyrolles.

Park, G., and G. Raskutti. 2015. “Learning Large-Scale Poisson DAG Models Based on Over Dispersion Scoring.” In *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. D. Lawrence, [D. D. Lee](https://papers.nips.cc/author/daniel-d-lee-1464), M. Sugiyama, and R. Garnett, 631–39. No place: Neural Information Processing Systems Foundation.

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

———. 1988. Probabilistic Reasoning in Intelligent Systems. Los Angeles: Morgan Kaufmann.

Rodrigues PP, Ferreira-Santos D, Silva A, Polónia J, Ribeiro-Vaz I. Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network. Artif Intell Med. 2018 Sep; 91: 12-22

Schwarz, G. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 1978: 6.

Shen, J., L. Li, and W. Wong. 2008. “Markov Blanket Feature Selection for Support Vector Machines.” In *Proceedings of the 23rd National Conference on Artificial Intelligence*, vol. 2, edited by Anthony Cohn, 696–701. Palo Alto, CA: AAAI Press.

Shojaie, A., and G. Michaildis. 2009. “Analysis of Gene Sets Based on the Underlying Regulatory Network.” *Journal of Computational Biology* 16: 407–26.

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. Berlin, Germany: Springer-Verlag.

Tan, Y., and L. Zhifa. 2013. “Feature Selection and Prediction with a Markov Blanket Structure Learning Algorithm.” *BMC Bioinformatics* 14 (17): A3.

Théophile H, Arimone Y, Miremont-Salamé G, Moore N, Fourrier-Réglat A, Haramburu F, Bégaud B. Comparison of three methods (consensual expert judgement, algorithmic and probabilistic approaches) of causality assessment of adverse drug reactions: an assessment using reports made to a French pharmacovigilance centre. Drug Saf. 2010 Nov 1; 33(11): 1045-54.

Vang, J. 2008. “Using a Model of Human Cognition of Causality to Orient Arcs in Structural Learning of Bayesian Networks.” PhD diss., George Mason University.

Verma, T., and J. Pearl. 1990. Equivalence and synthesis of causal models. UCLA, Technical Report, 1990.

Zargoush, M, F. Alemi, V. E. Vinzi, J. Vang, and R. Kheirbek. 2014. “A Psychological Approach to Learning Causal Networks.” *Health Care Management Science* 17 (2): 194–201.

Zeng, Y., L. Jian, and S. Lin. 2009. “Classification Using Markov Blanket for Feature Selection.” In *2009 IEEE International Conference on Granular Computing*, edited by T. Y. Lin, X. Hu, J. Xia, T.­-P. Hong, Z. Shi, J. Han, S. Tsumoto, and X. Shen, 743–47. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

1. [↑](#endnote-ref-1)
2. [↑](#endnote-ref-2)
3. [↑](#endnote-ref-3)
4. [↑](#endnote-ref-4)
5. [↑](#endnote-ref-5)
6. [↑](#endnote-ref-6)
7. [↑](#endnote-ref-7)
8. [↑](#endnote-ref-8)
9. [↑](#endnote-ref-9)
10. [↑](#endnote-ref-10)
11. [↑](#endnote-ref-11)