Causality, Confounding, and Simpson's Paradox

Asad Zaman[1®] and Taseer Salahuddin

Received: 11.01.2020 Accepted: 04.04.2020 Published: 30.04.2020

ABSTRACT

This is an introductory article which explains the importance of explicit consideration and modeling of causality, contrary to current econometric practice, in order to use data set for extraction of meaningful information. One of the easiest to understand approaches to causality is via Simpson's paradox. We will use this paradox, framed in different real-world contexts, to provide an introduction to basic concepts of causality.

Key words: *Simpson's Paradox, Causality, Econometrics, Confounders, Mediators*
JEL Classifications: C0, C1

1. INTRODUCTION

Bitter fighting among Christian factions and immoral behavior among Church leaders led to a transition to secular thought in Europe (see Zaman, 2018, for details). One of the consequences of rejection of religion was the rejection of all unobservables. Empiricists like David Hume rejected all knowledge which was not based on observations and logic. He famously stated that: "If we take in our hand any volume; of divinity or school metaphysics; for instance, let us ask, does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames: for it can contain nothing but sophistry and illusion." David Hume further realized that causality was not observable. This means that it is observable that event *Y* happened after event *X*, but it is not observable that *Y* happened due to *X*. The underlying mechanisms which connect *X* to *Y* are not observable. In the early 20th Century, the philosophy of logical positivism which says that human knowledge is based solely on observations and logic became widely accepted and wildly popular. Disciplines like Statistics and Econometrics, which evolved during 20th century, were built on positivist foundations. They only deal with measurable and observable (numbers) and ignore immeasurable concepts like causality. A much more detailed discussion of the philosophical background which led to these widespread misconceptions about human knowledge is given in Zaman (2012).

Pearl et. al. (2016, Chapter 2) provide a history of how mistakes by the founders of the discipline led to replacement of causality by correlation in statistics. Pearl and Mackenzie (2018, Chapter 5) provides the history of how causal information was dropped from econometric models. Current econometric techniques do not allow us to distinguish between real and spurious relationships. Excellent and robust fits can be seen between totally unrelated variables like log of number of newspapers published and life expectancy; there is no way to tell if a regression

® Asad Zaman, Professor of Economics, Former Vice Chancellor, Pakistan Institute of Development Economics, Islamabad, Pakistan (email: asad.zaman@alumni.stanford.edu), Tel: +92 334 5164905.
Taseer Salahuddin, Assistant Professor of Economics, Government Sadiq College Women University, Bahawalpur, Pakistan (email: salahuddin.taseer@gmail.com), Tel: +92 3036660871

is real or spurious. How do we differentiate between a regression of Turkish Consumption on Turkish GDP, which has a strong causal basis, and one of GNP on Newspapers, which is purely correlation without causation? Many examples like these are discussed in Zaman (2010), which show serious confusions about causality in conventional econometrics, and resulting consequences in terms of defective analysis.

This is an introductory article which explains the importance of explicit consideration and modeling of causality, contrary to current econometric practice, in order to use data set for extraction of meaningful information. One of the easiest to understand approaches to causality is via Simpson's paradox. We will use this paradox, framed in different real-world contexts, to provide an introduction to basic concepts of causality.

2. THE BERKELEY ADMISSION CASE

Suppose there are two departments Engineering (ENG) and Humanities (HUM), which have differing admissions policies. Due to these policies, 80% of female applicants to ENG are admitted, while only 40% of the female applicants are admitted in HUM. To understand Simpson's Paradox, it is essential to understand the relation between these departmental admissions rates, and the overall admit rate for females in Berkeley. Assuming, for simplicity, that these are the only two departments, we ask: What is the OVERALL admission rate for female applicants at Berkeley? The answer is that the overall admit ratio is the weighted average of the two admission percentages (80% and 40%). Table 1.1 shows overall admit rate of females with different number of applicants.

	Engineering			**Humanities**			**Overall admit rate**		
Situations	Applied	Admitted	% Admitted	Applied	Admitted	% Admitted	Applied	Admitted	% Admitted
A	1800	1440	80%	200	80	40%	2000	1520	76%
B	1500	1200	80%	500	200	40%	2000	1400	70%
C	1000	800	80%	1000	400	40%	2000	1200	60%
D	500	400	80%	1500	600	40%	2000	1000	50%
E	200	160	80%	1800	720	40%	2000	880	44%

Table 1.1 Overall admit rate for Female applicants.

If all females apply to HUM and none to ENG then overall admit rate is 40%. If all females apply to ENG then overall admit rate for females will be 80%. The table shows that the overall admit rate for females can vary from 40% to 80% depending upon proportions of females which apply to the two departments.

Now suppose Berkeley systematically discriminates against males. For male applicants to ENG, the admit ratio is only 60%, much lower than the 80% ratio for females. For male applicants to HUM, the admit ratio is only 20%, much lower than the 40% for females. What will the overall admit rate for males be? As before, this will be a weighted average of the two rates 20% and 60%, where the weights will be the proportion of male applicants to the two departments. The table below shows how the overall admissions ratio varies depending on how many males apply to which department:

	Engineering			Humanities			Overall admit rate		
Situations	Applied	Admitted	% Admitted	Applied	Admitted	% Admitted	Applied	Admitted	% Admitted
A	1800	1080	60%	200	40	20%	2000	1120	56%
B	1500	900	60%	500	100	20%	2000	1000	50%
C	1000	600	60%	1000	200	20%	2000	800	40%
D	500	300	60%	1500	300	20%	2000	600	30%
E	200	120	60%	1800	360	20%	2000	480	24%

Table 1.2 Overall admit rate for male applicants.

The table shows that the overall admit rate for males can vary between 20% and 60% according to how the applicants are distributed between ENG and HUM. We have already seen that overall admit rates for females can vary between 40% and 80%. Now consider the scenario created by the highlighted rows in the table. If 90% of the females apply to HUM, then the female admit ratio will be 44%, close to the 40% admit ratio for females in HUM. If 90% of the males apply to ENG then the admit ratio for males will be 56%, close to the 60% admit ratio for males in ENG. Despite the fact that females are heavily favored in both ENG and in HUM, the overall admit ratio for females (44%) will be much lower than the admit ratio for males (56%). Someone who looks only at the overall admit ratio for males and females will come to the conclusion that Berkeley discriminates against females, which is the opposite of the picture that emerges when looking at departmental admit ratios. This is known as the Simpson's Paradox. Interestingly, this is not a hypothetical example. I have simplified the numbers to make the analysis easier to follow, but the actual data for Berkeley admissions follows a similar pattern. The overall admit rates appear to show bias against females. Bickel et. al. (1975) carry out a standard statistical analysis of aggregate admissions data. They test the hypothesis of equality of admit rates for males and females and conclude that males have significantly higher admissions ratio than females. A causal analysis of data attempts to answer the "WHY" question. Why is the admit rate for males higher? To try to learn why the male admit rate was higher, Bickel et. al. (1975) looked at the breakdown by department. Note that the data themselves furnish us with no clue as to what else we need to look at. It is our real-world knowledge about colleges, admissions process, departments, which suggests that department-wise analysis might lead to deeper insights. This shows how real-world knowledge, which goes beyond the data, matters for data analysis. Doing the analysis on the departmental level leads to an unexpected finding – each department discriminates in favor of women. Philosophers call this "counter-phenomenal". The phenomena – the observation – at the aggregate level suggests that Berkeley discriminates against women. But a deeper probe into reality reveals that the opposite is true. This shows the necessity of going beyond the surface appearances, the observations, to deeper structures of reality, in order to understand the phenomena. This is in conflict with Empiricist ideas that observations by themselves are sufficient, and we do not need to probe deeper. See Zaman (2020, Models and Reality, this issue) for a more detailed discussion of the philosophical background.

When we discover a conflict between the phenomena and our exploration of the phenomena – the deeper and hidden structures of reality – then we are faced with the necessity of explaining this conflict. Because both departments discriminate against males, the explanation that

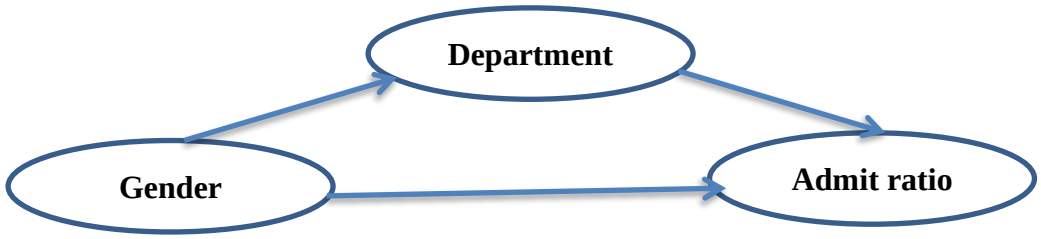
Berkeley admissions process discriminates against females is no longer acceptable. Bickel et. al. (1975) do the data analysis and come up with the deeper explanation. ENG is easier to get into, and HUM is more difficult. Females choose to apply to the more difficult department and hence end up with lower admit ratios. Males choose to apply to the easier department, and hence have higher admit ratios. The search for causal explanations does not stop here. We can then ask: WHY do females choose humanities? We can also ask: WHY is ENG easier to get into, and WHY is HUM more difficult to get into? For both of these questions, there are several possible hypotheses which could be true, and which could be explored using data or qualitative techniques. In the next section, we consider some other causal structures for admissions, which lead to radically different answers to the WHY questions, even though the observed data remains exactly the same.

Exercise 1: Even though causality concepts are relatively easy, lack of familiarity leads to difficulties in grasping them. For this purpose, it is useful to grapple with them directly via exercises. Suppose that there are two main hospitals A and B in a small city. For patients admitted to hospital A, recovery rates are 44% - less than half recover. On the other hand, 56% of the patients admitted to hospital B recover. On the basis of these numbers, it appears that hospital B is the better of the two, with higher recovery rates. Think of a hidden factor which might reverse this comparison. Create an example of Simpson's Paradox by making up data categorized along the hidden factor which shows the paradox.

2.1. Distinguishing Between Confounders and Mediators

Before the impact of changing causal structures on analysis can be understood two different points need to be made. Firstly, as noted by Pearl and Mackenzie (2018) and Hoover (2004), a major obstacle to development of causal thinking has been the lack of mathematical notation to express causality. In particular, the "equal sign" in a regression model ($C=a+bY$) does not give us a clue that Y is a cause of C, and encourages the mistaken algebra that $Y=(C/b)-(a/b)$, which is not correct as a causal equation. One of the critical factors in the development of understanding about causality has been the development of a natural language to express causal relationships. This is the language of path diagrams, which we now introduce. The causal structure of the explanation given by Bickel et. al. (1975) for Simpson's Paradox in Berkeley admissions can be depicted in the following path diagram (Figure 2.1).

Figure 2.1 Causal Map # 1A (Bickel)



Here gender affects both the choice of department and admit ratio. Department also affects admit ratio. The causal path diagram is crucial in determining exogeneity and endogeneity. These are key concepts used in Econometrics, but they remain confusing because they cannot be defined correctly without understanding causality. Path diagrams are required to distinguish between the two cases. Furthermore, as we will see later, this classification is not sufficient for analysis. Among endogenous variables we must distinguish between mediators and colliders, as the two cases require different analysis. In this section, we will explain the difference

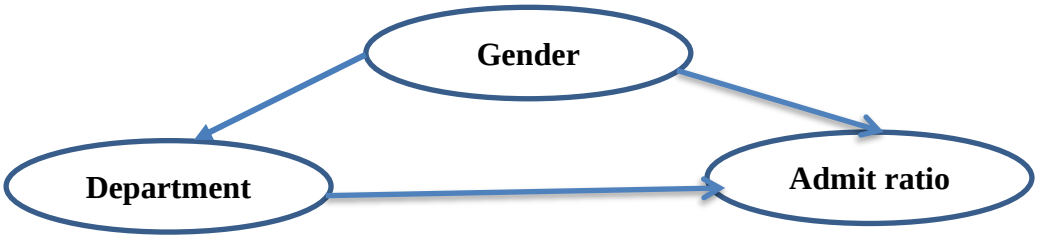
between exogenous variables and endogenous variables within the context of the Berkeley Admissions case under discussion.

When we are studying the effect of gender on admissions, the above diagram shows that there are two causal pathways. One is the direct effect of gender on admissions. This effect leads to a 20% increase in chances of admissions over the admissions rate for males. The second is the indirect affect which operates via the choice of department. ENG is easier to get into, and HUM is harder. If gender influences choice of department, so that females choose the harder department while male choose the easier one, then this will create an adverse impact on admissions. The total effect will be the sum of the two effects, direct and indirect. In this causal path diagram, the department is called the mediator; it acts a mediator of the effect of gender on admissions.

In terms of econometrics, think about the Admit Ratio as the variable to be explained, and Gender and Department as the explanatory variables. Modeling admissions via a function like $AR=f(G,D,G\text{x}D)$ will not correctly capture the causal relationships. The coefficients of Gender, Department, and interaction terms will not correctly represent the impacts of the three variables on the admissions rate, because these do not take into account the one-way causal path between Gender and Department. Structural Equation Models (SEM) can capture such relationships and potentially lead to the right analysis, although current practice does not utilize SEM for causal modeling.

In this same diagram, we can also study how admissions policies vary across departments. Gender affects both Department and Admissions. The causal path diagram is the same as the previous one, but the question of interest is now different

Figure 2.2 Causal Map # 1B (Dept → Admissions)



The path diagram shows that Gender influences both Department choice and the admit ratio, but it is NOT influenced by either of these variables. This makes Gender exogenous. To be more precise: When we are studying the effect of department on the admission ratio, gender is exogenous because it is not causally influenced by either of the two variables under study. Statisticians say that Gender is a confounding variable, but are equally confused about the meaning of confounding, because understanding it requires causality. To solve this problem of confounding, we must CONDITION on gender. In the language of econometrics, exogeneity means that we can condition on this variable, treating it as constant. Holding gender constant -- i.e. separately calculating admit rates for both genders – allows us to calculate the effect of department choice on admit ratio, while preventing the confounding variable from changing, so as to eliminate its effects.

Exercise 2: Consider the hospital example of exercise one. Suppose there are two types of patients – critical and non-critical. This has a direct impact on recovery rates, with critical patients having lower recovery rates. Suppose hospital A is better equipped to deal with critical

patients, so whenever possible, critical patients choose to go to hospital A. Recovery rates are different at hospitals A and B, and hospital A is better at dealing with both kinds of patients. Nonetheless, overall recovery rates are higher at hospital B. Answer the following questions

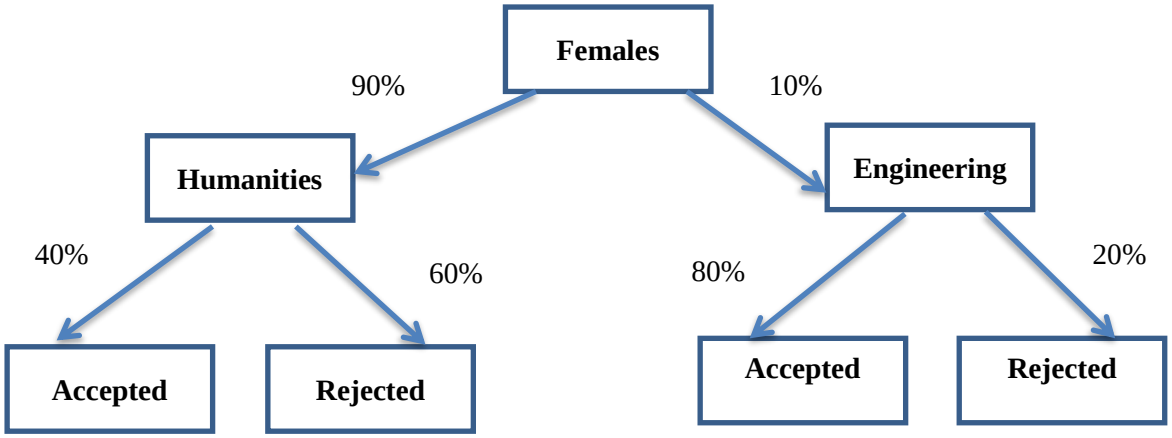
1. Draw a causal path diagram connecting Patients (critical and non-critical) to recovery rates. According to your diagram, is the hospital a mediator or a confounder? Give a verbal explanation of the paradox using the conflict two factors which affect recovery rates, or the two causal pathways.

2. Suppose we wish to evaluate the performance of the hospitals. Draw a diagram connecting Hospital to Recovery Rates. Is the type of "patient" a mediator or a confounder in this diagram? Explain how ignoring the confounder will lead to misleading comparison of hospital performance. Also explain how conditioning on the confounder leads to a correct analysis.

2.2. Calculating Probabilities on Decision Tree

Consider the problem of calculating the effects of Gender (G) and Department (D) on Admissions Rates (AR). In regression terminology, both G and D are exogenous regressors, while the dependent variable is AR. In regression methodology, there is no way to express the fact that G influences D. Furthermore, the observed data distributions remain the same in the three cases where $G \rightarrow D$, $D \rightarrow G$, and where both are independent. In the causal case under consideration, where Gender influences choice of Department, calculating the effect of Gender on admissions rate requires the use of a decision tree, as we now illustrate. On the other hand, if we want to calculate the effect of department on admission rate, standard regression methodology works well. In effect, regression methodology conditions on gender, and calculates the partial effect of department on admission while holding gender constant. The tabled numbers, which show 80% and 60% admissions rates for females and males in ENG, and 40% and 20% in HUM, correctly display the impact of department on admissions holding gender constant. On the other hand, these tables do not provide the right picture when we want to calculate the effect of gender on the admissions rate. To calculate the impact of gender on admissions, we must take into account both channels by which gender affects admissions. This can be done conveniently via a decision tree diagram, as we now illustrate. Say for example John and Jane apply to Berkeley. The following tree diagrams show how we can calculate the probabilities of admissions for the two via separate tree diagrams.

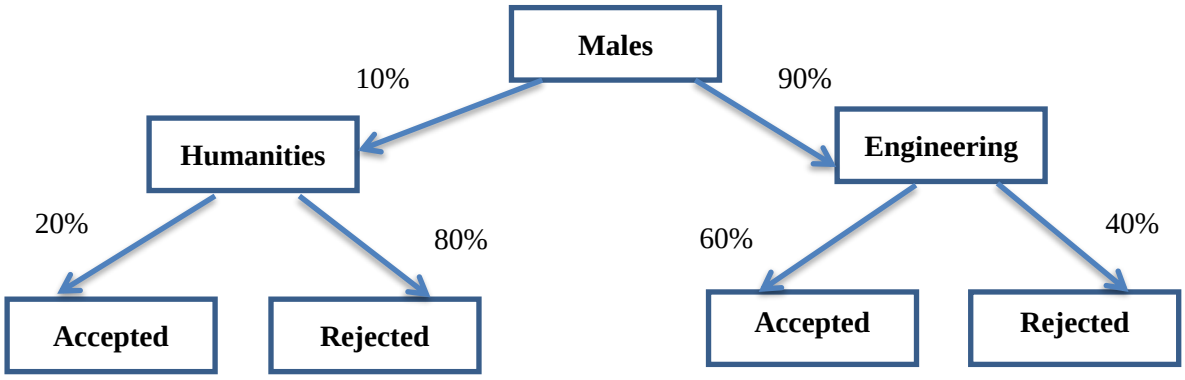
Figure 2.3 Female Admit Probabilities Tree Diagram



Only 10% females choose Engineering, where they have 80% chances of admission and 90% females choose Humanities but with only 40% chances of admission. Applying this past data, we can estimate that Jane will apply to HUM with 90% probability and to ENG with 10% probability. Using the admit rates for females in ENG and HUM, we can compute her chances of getting admission as: 90% *40% + 10%*80% = 36% + 8% = 44%. This calculation is valid under the assumption that the past history of admissions applies to Jane – she is not differentiated from the typical female applicant in ways which would make the history irrelevant.

For John, we have the same diagram, but the probabilities are different at each branch of the tree for males.

Figure 2.4 Male Admit Probabilities Tree Diagram



John has a 10% chance of applying to HUM and a 90% chance of applying to ENG. In the two departments, his admit probabilities are 20% and 60% respectively, so his overall admit probability is 10% *20% + 90%*60% = 2% + 54% = 56%

These calculations show that the probability for females to get into Berkeley is only 44%, while the probability for males is 56%. These aggregate numbers correctly reflect overall admit probabilities categorized by Gender (without conditioning on the Department). On the other hand, if the department choices for Jane and John are known, then entirely different probabilities would apply. The question of when we should condition, and when we should not, depends entirely on the research question we are trying to answer. There is considerable confusion about this matter mainly because there is no way to explicitly take into account causal information in conventional statistical and econometric analysis.

Exercise 3: Patients (critical and non-critical) choose hospitals, and this choice of hospitals determines the recovery rates. Draw separate tree diagrams for critical and non-critical patients showing probabilities of choice of hospitals, and recovery rates at each hospital, to get the overall recovery rates for both types.

2.3. Changing Causal Structures

In the above given discussion, there is not one (gender) but two causal factors (gender and difficulty level per department) that influence the admission rates. Being female is a positive influencer in each department but applying to humanities is a strong negative influencer. As men applied more to engineering their negative influencer (gender) was overwhelmed by their positive influence of department choice (engineering). Bickel et. al. (1975) noted this paradox, and provided this explanation. This explanation is just a conjecture about the causal factors

which affect admissions rates. It is possible that taking into account other causal factors (like GPA) could lead to entirely different analysis of exactly the same data. For example, suppose admissions policies at ENG and HUM are only based on GPA, but HUM is more difficult to get into:

	Admit Percent	
Grade	ENG	HUM
A	90%	60%
B	70%	30%
C	50%	10%

If female applicant pool has better grades (i.e. 60% A's, 30% B's and 10% C's) while male applicant pool has significantly lower GPA's (10% A's, 30% B's and 60% C's), then exactly same admission patterns will be observed. The original data appears to show gender bias – females are preferred. But deeper analysis reveals that there is no discrimination by gender. The admissions criteria are the same for both genders, but more females get in because they have higher grades. Now the causal question here would be why better male applicants do not apply to Berkeley while better female applicants do? Answer may lie in availability of a good all-male university which is not available for females. Or, one could construct examples where females are better at English and males are better at math, and this provides the causal explanation. Now these answers to the WHY questions do not come from numbers; we are speculating about unobserved, real structural factors. This is the central point of the article: numbers do not contain causal information, whereas understanding the numbers needs causal understanding, which only comes from qualitative kinds of knowledge about deeper structures of reality. Current econometric practices are deeply flawed because they can only be used to describe correlations among observations, and do not take causal information into consideration. Even worse, they encourage intellectual laziness among students, who are taught to believe that running a regression yields reliable results without any considerations of causal structures.

Exercise # 4: The discussion above sketches the possibility that admissions depend solely on GPA, and female applicants have higher GPA, which is what accounts for apparent preference for females. Create the statistics to go with this story, and complete the sketch of the data provided in the first paragraph.

2.4. Policy Depends Upon Unobservable Causal Relations

As Hume realized a long time ago, observations cannot reveal the underlying causes. Studying alternative hidden causal structures for the same data set could radically change the interpretations discussed in previous sections. Understanding data REQUIRES understanding causal structures which generate the data and these causal structures are not DIRECTLY observable. Unfortunately, a strong intellectual tradition in the West supports a deeply mistaken understanding of scientific methodology. According to the empiricists, the surface appearances, and the observations, are all that we have. We cannot go beyond them, and we should not go beyond them to explore the causal structures of hidden reality. See Zaman (2012) and Zaman (2020, this issue) for a more detailed discussion of this philosophical background. As we have seen, taking causal structures into account can show us that the surface appearances are deceptive. What appears to be discrimination against females is actually caused by females choosing the more difficult department. This point can be understood more clearly by further simplifying the causal structure. Suppose admission in departments is gender blind. Engineering gives 60% admission to all applicants, male and female. Similarly, Humanities

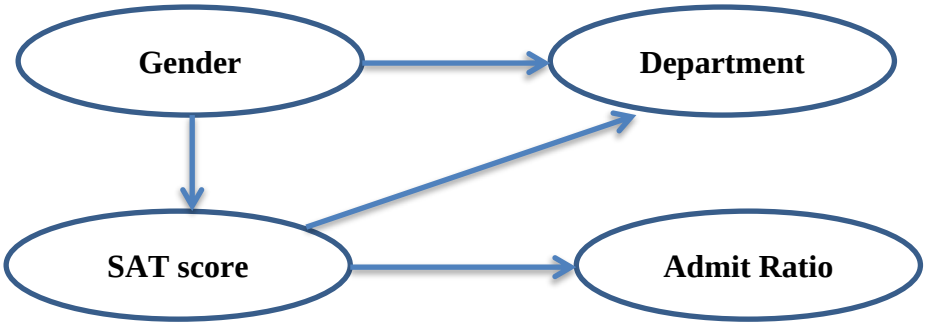
admits 30% of all applicants, male and female. There is absolutely no difference in admissions rate by gender in either department.

Figure 2.5 Causal Map # 2

Even though gender has no direct impact on the admit ratio, gender does affect department choice, and department choice affects the admit ratio. Because of this causal chain, gender will appear to affect the admit ratio. As an extreme example, if all females apply to Humanities and all males to Engineering, then female admit ratio will be 30% and male admit ratio will be 60%, providing strong evidence of apparent discrimination against females. This shows the importance of the studying the MECHANISM. The data reveal that gender affects the admit ratio, but they do not tell us HOW this happens. To study the causal chain, we must follow the real-world process closely; in Freedman's (1991) terminology, we must expend shoe leather to learn about causes. As we follow the admissions process step-by-step, we will see that females are mostly applying to HUM while males are applying to ENG. This would lead us to ask whether the different admit ratio is due to different policies at the two departments, rather than gender. The data can shed light on this question, and lead us to a rather different answer to the WHY question. Females have lower admit rates because they apply mostly to HUM which is more difficult to get into. This raises further WHY questions – why is HUM more difficult to get into? Perhaps it is because of limited funding which leads to a smaller department size and hence more competition for fewer seats. Or many other factors could be causes. Searching for causes requires looking at structures of external reality beyond what appears on the surface. Causes do not reveal themselves in the observed statistics.

To drive home this point, we consider other causal structures, which can create outcomes resembling original ones but with causes entirely different from those discussed above. Suppose admission process is totally computerized and only dependent upon SAT scores. Suppose SAT scores *X* leads to an *X*% chance of admission. As an artificial example, suppose female SAT scores are only 80 and 40 while male SAT scores are only 60 and 20. Suppose all females with 80 SAT score apply to Engineering, while females with 40 SAT score apply to Humanities. Similarly, all males with 60 SAT score apply to Engineering and those with 20 apply to Humanities. While these numbers are cooked to make this easy to understand, it is easy to construct more complex realistic examples which would create the same phenomena.

Figure 2.6 Causal Map # 3



Using initial example proportions 200 females with 80 score and 1800 males with 60 scores apply to Engineering while 1800 females with 40 score and 200 males with 20 score apply to Humanities. This will create exactly the same results as Table 1.1, even though gender and department have no effect on admissions as admissions are purely based on SAT scores.

The causal diagram portrays our assumptions about the causal structure. Each arrow represents one of our causal assumptions. We have assumed that Gender affects SAT scores as females have higher SAT scores. We have also assumed that Gender also affects department choice as females overwhelmingly apply to Humanities. The SAT score also affects department choice as males and females with high SAT opt for Engineering and with low SAT score opt for Humanities. Finally, the admissions process depends ONLY on the SAT score; it does not depend on department, and it does not depend on Gender. But the observed data from this assumed causal structure will be exactly the same as in Table 1.1 showing Simpson's paradox. An analysis which ignores the causal structure, will come to several wrong conclusions. The data will show that

A. Gender and department both affect admission rates.

B. Females are strongly preferred by both departments.

C. Engineering has higher admit rates and is easy to get into. Humanities has lower admit rates and is more difficult to get into.

The assumed causal structure makes all three obvious and strong statistical relationship turn out to be false. Gender does not affect admit ratios, only SAT scores do. The higher admit rates for females at ENG and HUM result from higher SAT scores of the females, and not from preference of females by the departments. The lower admit rates in HUM are not because it is more difficult to get into HUM. It is because males and females with lower SAT scores apply to HUM – both departments have exactly the same admit policies. If the causal structure is as hypothesized, then an analysis using SAT scores can be used to prove these *counter-phenomenal* claims – claims which contradict the surface appearances. All of the apparent relationships A, B, and C, will disappear when we condition admit ratios on the SAT Scores. Once we do this, we will see that there is no effect of gender and no effect of department on admit ratios.

This analysis has disturbing implications. Given data on Gender, Department, and Admit Ratios, the statistics strongly supports A, B, and C. How can we learn that this analysis is wrong? There is NOTHING in the data which will give us a clue. If we have expended some shoe leather walking around Berkeley, talking to faculty, learning about the admissions process, then we could come up with a causal structure. For example, if we learn that admission occurs by a mechanical and computerized process which only looks at SAT Scores, then we would realize that our hypothesized causal relationships represented by A, B, and C are wrong. Further study could reveal the correct causal structure, leading us to gather data on SAT Scores, and then to the right conclusions. Suppose on the other hand that the data about gender, department and admissions comes from some remote and unknown university in Siberia, about we have no further information, and no possibility of expending shoe leather to find out more. Then we could tentatively assert that the data suggest conclusions A, B, and C, but we cannot be sure, because we do not know the causal mechanisms.

Understanding causality is especially important for big data and machine learning, because computers cannot expend shoe leather. Causal relationships are learnt from real world experiences and knowledge. For example, if we know about how the applications process works at Berkeley, we could rule out some causal sequences, and also assess plausibility of others. But an abundance of data would not lead us to knowledge of these structures, and machines could not discover those structures, because they are not present in the data.

Exercise 5: Continue from exercise 4. Create an example where the data, when analyzed by gender and department, show that females are preferred by both departments, and that HUM is more difficult to get into, while ENG is easier. Now introduce the hidden variable GPA, and make it do the work of the SAT Score. Once we use GPA, the initial findings are all reversed. We find that admissions policies are exactly the same at ENG and HUM – neither department is easy or difficult. We find that admission is gender blind – same policy for males and females. Explain why it appears that females are favored, and why it appears that HUM is more difficult to get into, even though this is not really the case

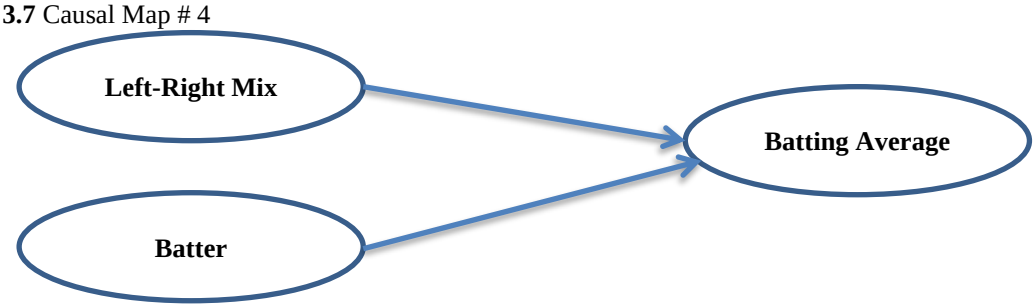
3. BASEBALL SCORES: OVERALL AVERAGE OR STRATIFIED?

It is standard practice in statistical analysis to separate the role of the field expert from the statistical consultant. The field expert has deep knowledge of the processes which generate the numbers, while the statistician knows about the numbers, and has superficial knowledge of the meaning of these numbers and where they come from. The assumption is that we can analyze data without knowing about the background mechanisms which generate the data. To show that this assumption is false, we will consider the same data of Berkeley admissions in Table 1.1 and give it very different real-world background. Suppose now that Tom and Frank are two batters with batting averages of 56% and 44%. On the basis of only these numbers with no knowledge of the field it appears that Tom is apparently the better one out of two. At a critical moment for the team, coach needs to decide whom to send into the field for getting a hit. From the overall averages, Tom seems to be the better choice. However, an analyst has heard that Frank is a superstar against left-handed pitchers, but has difficulties when he faces right-handed pitchers. Also, generally Frank has a much better reputation than Tom. Thus, the superiority of Tom's average puzzles the analyst. On the basis of real-world information, available to field experts, but not to statisticians, the analyst decides to look deeper into the batting records. Analyzing performance against left and right-handed pitchers, he finds that Frank had 80% hits against left-handed pitchers while Tom had 60%. Also, Frank had 40% hits against right-handed pitchers whereas Tom had only 20%. Thus, Frank is better than Tom against left-handed Pitchers, and he is also better than Tom against right-handed pitchers. On the basis of this deeper analysis, it seems that the Coach should send out Frank. This is because digging deeper into the data provides a counter-phenomenal answer to the question of WHY did Tom (56%) have a higher batting average than Frank (44%). The obvious answer would be that Tom is better than Frank. The subdivided data shows that this is false. Frank is better than Tom, but both batters perform better against left-handed pitchers, and worse against right-handed pitchers. So the CAUSE of Frank's poorer performance is that he faced far more left handed pitchers where his batting average is 40%, while Tom faced far more right-handed pitchers where his batting average is 60%. Given the same mix of pitchers, Frank would perform much better than Tom. This analysis is captured by the following causal diagram. The batting average depends on which batter, but it also depends on the proportion of left and right-handed pitchers.

According to the causal map, the left-right mix of pitchers is not caused by the batter or the batting average, and hence is exogenous. In statistical terminology, it is a confounder. If we

want to correctly calculate batter performance, then we must condition on the confounder. This means holding left and right-pitcher constant, and calculating separately for the two cases. One the basis of this calculation, we would send out Frank, because he will perform better than Tom against any arbitrarily chosen left/right mix of pitchers. Here the separate performance numbers against left and right-handed pitchers are relevant, and the aggregate performance is misleading because the two players did not face similar left/right mix. However, a different causal structure can lead to the opposite conclusion, as we now discuss and demonstrate

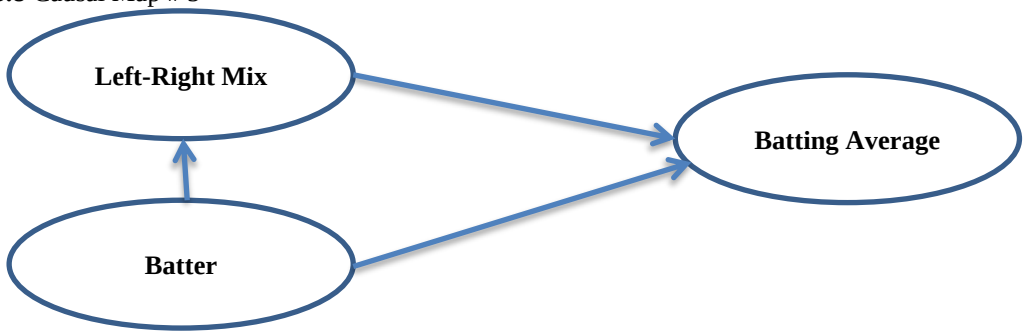
Figure 3.7 Causal Map # 4



3.1. Coach is Aware of Exceptional Performance of Tom

Expending shoe leather, our analyst goes to have a talk with the coach of the opposing team. The coach informs him that Frank is a superstar, and has an amazing batting average of 80% against left-handed pitchers. However, his performance against right-handed pitchers is not so good, and so, whenever he is sent to bat, the coach switches to right-handed pitchers as much as possible. With this piece of causal information, which simply cannot be contained in any kind of data, the causal picture changes to the following:

Figure 3.8 Causal Map # 5



Now the mix of left and right-handed pitchers is NOT exogenous, because the coach changes the mix in response to Frank. When the mix is endogenous, then the batter's influence on batting average works through two channels. One channel favors Frank – he is the better batter of the two. The other channel has a negative influence – the coach changes the pitcher mix to make it difficult for Frank. The overall effect is the sum of the two effects and is negative in the present example. It is better to send Tom, even though he is weaker, if he will get a more favorable left/right mix. So, depending on the causal chain, either Tom or Frank could be the better choice. The causal chain cannot be learned from the data, but we can learn about it by expending shoe leather and examining the real-world mechanisms in operation which create the numbers we see. Some important points that we can learn from these examples are as follows:

1. The causal chain depends crucially on the unobservable intentions of the coach of the opposing team. Will he respond to Frank by changing the mix of left/right batters? Or is his choice of mix independent of which batter is chosen?

2. Even though this factor is unobservable, and we can never have 100% certainty about it, we can still make intelligent guesses about it. So, even when causal factors can never be known with certainty, we can make good and bad guesses about them, and the results we get will depend on how well we can guess at unobservables.

Many philosophers (Hume, Kant, Wittgenstein, Logical Positivists) have maintained that we should not speculate about unobservables, and that unobservables should not enter scientific theories. The above example shows that, to the contrary, we must speculate about unobservables, as they matter crucially for our decisions. Another widely accepted philosophical principle is that use of the term "Knowledge" should be confined to those matters about which we can be certain. Again, all of our examples show that we can only have guesses about causal mechanisms, and we can never be certain that some new hidden variable or causal effect may completely upset all our previous calculations. But despite this, we have no choice but to try to do our best at making such guesses, without ever being certain that we are right. Meaningful analysis cannot be done without causal structures, and yet, we can never be certain that we have correctly discovered the causal structure. Wrong philosophies about the nature of science and knowledge have prevented understanding of causality in statistics and econometrics until recently.

Causal analysis cannot be applied mechanically. In making the decision about which batter to send out, we must assess the relevance of the historical record to the current situation. Are the observed batting averages and Left and Right-Handed pitchers reasonable estimates for performance against the current field? The analysis that the overall average is better predictor of performance, and Tom should be sent out, depends on several assumptions. We assume that there is a normal standard left and right pitcher mix which has been chosen by the coach for general use. For any given fixed left/right mix, Frank will do better than Tom. However, if the Coach switches the mix in response to Frank – his exceptional performance is the terror of opponents – then Frank will face an adverse mix. We also assume that Tom does not stand out sufficiently from others, and the Coach will not optimize pitchers against him. With these assumptions, all about unobservable causal chains, it would make sense to send Tom out to bat, even though Frank is the better batter among the two. Changes in the mind of coach would change causal structures, and change optimal decisions, so we must make guesses about this. For example, if the coach decides to optimize pitcher against all incoming batters, then Frank would be a better choice – he would bat 40% against Tom's 20% for the left-handed pitcher to be chosen. If we have no direct clues about how the coach will choose, we may look at the historical record. This record shows that in the past, Frank has faced a highly adverse field, seemingly optimized against him, while Tom has not. So, there is a suggestion, a clue that coaches tend to optimize against Frank, but not against Tom. This would lead to choosing Tom.

Exercise # 6: Construct a decision tree as follows. Coach A selects either Tom or Frank. Then Coach B selects either a left-handed pitcher, or a right-handed pitcher. Use the probabilities PT and $PF=1-PT$ for Coach A, and PL and $PR=1-PL$ for Coach B to calculate the probability that Coach A gets a hit – use other data as given above. Consider the following three cases:

1. Both coaches choose their probability separately and independently. Show that regardless of how they choose, Tom is the better choice for coach A.

2. Coach A has a standard set of probabilities PL^*=50% and PR^*=50% which he generally uses to choose pitchers. However, when Coach A sees Tom at bat, he responds by changing to PL^*=10% and PR^*=90%. Show that in this case, Frank is the better choice.

3. Coach B knows the hit probabilities for Tom and Frank against left and right-handed pitchers, and chooses to optimize – that is, he chooses the type of pitcher against which the batter has worst performance. Show that in this case, Tom is the better choice.

4. EFFECT OF DRUGS ON RECOVERY

Our final example considers the classical case of testing the effectiveness of a drug as a treatment for a disease. It is standard to divide the population into two groups, and compare recovery rates. The group which does not take the drug is called the control group. The group which takes the drug is called the treatment group. If the percentage of people who recover from the disease in the two groups is roughly the same, then the drug has no effect on recovery. If the recovery rate in the treatment group is significantly higher than the recovery rate in the control group, then the drug is effective. It is also possible that the drug is actually harmful, in which case the recovery rate in the treatment group would be lower than that in the control group. All this is under the assumption that the treatment group and control group are sufficiently similar, so that the only difference between the two is that one group took the drug while the other one did not. If there are other significant differences between the groups, then the difference in outcomes could be due to the other factors, called confounders. For example, if the treatment group is mostly female while the control group is mostly male, then gender is a confounding factor. If the disease and drug act differently on different genders – for example, females are resistant to the disease and recover quickly with or without drug – then what we think is improved performance due to the drug would actually be improved performance due to the fact that there were more females in the treatment group. Thus, the effect of the drug is confounded with the effect of gender.

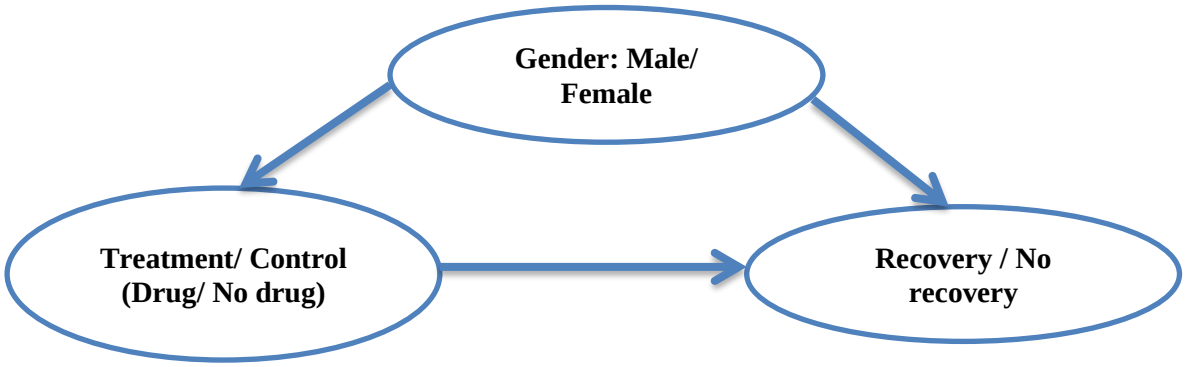
After establishing the basic terminology, we create an artificial data set parallel to Berkeley admissions, and show how different causal chains lead to radically different interpretations of exactly the same data. One source of confusion in this comparison is that in Berkeley, we were comparing the effect of gender on admissions, while here we are comparing the effect of drug treatment on recovery. So, being a female applicant to Berkeley corresponds to being in the treatment group, while being a male applicant to Berkeley corresponds to being in the control group. Also, in this comparison, the effect of gender corresponds to the effect of the department. The department ENG corresponds to being female while the department HUM corresponds to being male in the present setup. With this translation, we can now create a Simpson's Paradox for drug treatment as follows.

The overall recovery rate in the treatment group is 56%, while it is 44% in the control group. These figures appear to show that the drug is beneficial, and improves recovery rates by about 12%. However, expending shoe leather, we find a lot of complaints from patients about apparently harmful effects of the drug. On this basis, we decide to investigate further. Noting that recovery rates seem to differ by gender, we investigate recovery rates separately for males and females. This leads to a surprise. We find that females who did not take the drug (female controls) have recovery rates of 80%. However, females who take the drug (females in treatment group) have much lower recovery rates of 60%. The drug is quite harmful, lowering recovery rates by 20%. The drug is also harmful for males. The male control group has recovery

rates of 40%. However, the male treatment group, which takes the drug, has much lower recovery rates of only 20%. How can it be that the drug harms males, and it harms females, but it helps the general population? This is the Simpson's Paradox.

The recovery rates for the CONTROLS (non-drug-takers) are 80% for females (ENG) and 40% for males (HUM). The recovery rates for the TREATMENT GROUP (drug takers) are 60% for females (ENG) and 20% for males (HUM). If 90% of the controls are males while 10% are females, then the recovery rates for the non-drug-takers would be 10% x 80% + 90% x 40% = 8% + 36% = 44%. Similarly, if the treatment group is mostly (90%) females and has only 10% males, then the recovery rate in the treatment group would by 90% x 60% + 10% x 20% = 54% + 2% = 56%. The overall figures show that among those who took the drug, recovery rate was 56%, while it was only 44% among those who did not. Now the causal picture is rather different from the Berkeley admissions – it can be graphed as follows

Figure 4.9 Causal Map # 6



In the Berkeley admissions, we wanted to study the effect of gender on admissions. The department choices were affected by gender and so formed a second channel by which gender affected admit ratios. Here we want to study the effect of taking the drug on recovery. Taking the drug cannot affect gender. However, gender does affect drug taking since females overwhelmingly take the drug, while males overwhelmingly avoid the drug. Since females have higher recovery rates than males both with and without drugs, their heavy presence in the treatment group creates a bias towards recovery. Males have lower recovery rates with and without drugs, so their heavy presence in the control group creates a bias against recovery in the control group. Together these biases overwhelm the genuine negative effect of the drug. According to standard statistical analysis and terminology, gender is a confounding variable. Alternatively, in econometric terminology, gender is exogenous. The effects of drug on recovery are contaminated by the effects of gender, both on drug taking behavior and on recovery rates. To eliminate this, we must condition on gender and hold it constant. When we do that, we find the correct result that the drug is actually harmful to both females and males. The overall figure gives us the wrong information because it ignores the confounding variable of gender. The high rates of recovery in the treatment group come from the favorable effect of being female, and the high proportion of females in the treatment group. This favorable effect overwhelms the harmful effect of the drug on recovery. Several "Why?" questions arise in this context. Why the treatment group is so heavily populated by females and why is the control group so heavily populated by males? This means that the choice between treatment and control could not have been random, because random choice would automatically balance the number of males and females to roughly equal. Perhaps, because it was an experimental drug being tested, patients were offered a choice to take the drug or not. In this case females overwhelmingly chose to be treated, while males overwhelmingly chose not to take the drug,

creating a bias in favor of the drug. Why? There are many possible hypothesis and explanations which could be explored and tested here. For a discussion of how causes are discovered in real life examples, see Freedman (1991) on Statistical Models and Shoe Leather.

Exercise #7A: The standard method of avoiding confounders is to use randomization. Show that if people are assigned to treatment or control groups at random, then the chances of an imbalance like the one of the example – with 90% females in one group and only 10% in the other – is almost zero. To be specific, suppose 100 patients are assigned at random to the treatment group from the general population which is 50% female. Each patient has a 50% chance of being male or female. Use the Binomial distribution to calculate a range of numbers $(L=50-N, U=50+N)$ such that the $P(L<F<U)=99\%$. The number of females should lie within L and U with 99% probability. You will see that the number of females cannot deviate very far from the 50% proportion, so that extreme bias of the kind used in the example cannot arise in randomized controlled experiments.

Exercise #7B: There are many situations where randomized controlled trials are not possible. For example, if we want to study the effects of one year of smoking on cancer rates, explain why we cannot do randomized assignment to treatment and control groups. Similarly, if we have an experimental drug of unknown efficiency, possibly harmful, we cannot ethically randomize patients into treatment and control groups. Sketch a plausible causal scenario where, if an experimental drug is offered for treatment, women overwhelming prefer to take it leading to 90% females in the treatment group. At the same time, men overwhelming prefer not to take it, leading to 90% males in the control group. Note that discovering causality is always an exercise of the imagination – coming up with stories which explain the patterns we see at levels deeper than is revealed by the data themselves.

4.1. Blood Pressure as Exogenous Confounder

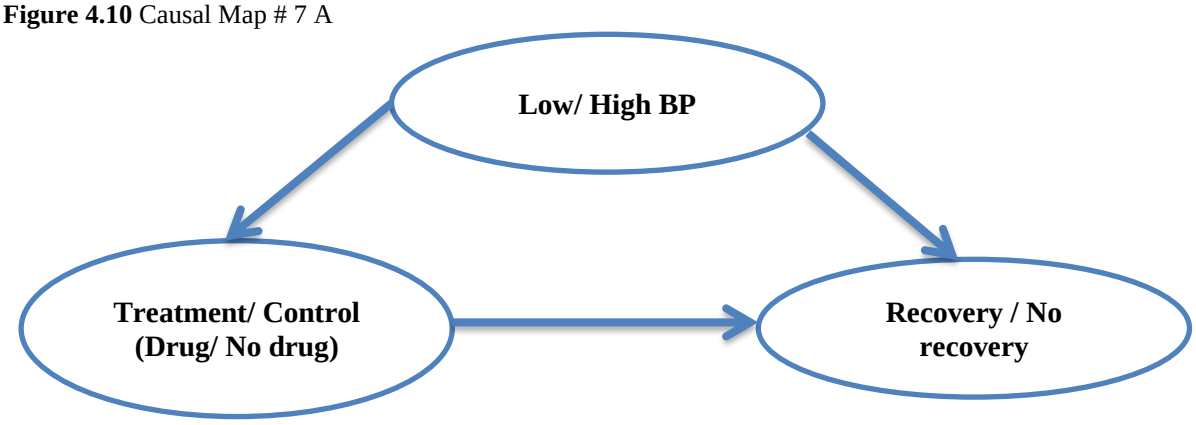
To see how changing causal sequencing changes the results completely, let us replace gender in previous causal map with Blood pressure BP. Gender is exogenous and cannot be affected by drugs and disease, but BP can be affected by disease and by drugs. This allows for a much larger variety of causal paths, each of which leads to different ways of understanding the same data set. Let us first consider the same causal structure as in the previous causal map #6, replacing Gender with Blood Pressure as a confounder. To keep the analysis parallel, suppose that there are only two types of BP: Low (LBP), and High (HBP). We will regard LBP as normal, for simplicity in interpretation. With this switch from Gender to Blood Pressure, the Simpson's Paradox takes the following form. Overall recovery rates are 56% for the treatment group and 44% for the control group, so that the drug appears to be beneficial. However, when we subdivide the population into LBP (normal BP) and HBP (elevated BP), we find the opposite results. In both subpopulations, the recovery rates of the control group are much higher than recovery rates in the treatment group, which suggests that the drug is actually harmful. To be more precise, in the HBP group, recovery rate is 80% in the control group (without the drug), and only 60% in the treatment group (with the drug). Similarly, in the LBP group, recovery rate is 40% in the control group, and only 20% in the treatment group. In both groups, the drug lowers the recovery rate by 20% and is obviously harmful. The paradox occurs because the control group consists mostly of LBP (90%) with only 10% HBP patients. LBP patients have low recovery rates of only 40% without the drug, so the presence of large numbers LBP patients drives recovery rates down to 44%. The treatment group has the opposite composition with 90% being HBP and 10% being LBP. HBP have 60% recovery rates (reduced from 80% after

being given the drug). After mixing in the 10% LBP patients with only 20% recovery rates, we get an average recovery rate of 56% in the treatment group.

How this data should be interpreted depends strongly on the causal chains, which are not available from the data itself. One possible interpretation is that the higher recovery rate is not due to the drug, which is harmful, but rather due to the large proportion of HBP patients, who have higher recovery rates even after being given the drug. This is the picture shown by recovery rates in the subgroups of low and high blood pressure patients. Which of the two statistics, overall or subgroup, gives the correct picture? Should we use the drug to treat patients, or should we discontinue its use? Is the drug harmful, or is it helpful? The answer to this very real and very important medical question depends on the hidden and unobservable causal chains which connect the three variables under consideration. Depending on the causal sequence either one of the two answers can be correct. Also, if there are other hidden factors, neither of the two answers may be correct. We first consider the simplest case, where blood pressure, like gender, is an exogenous confounder.

Case 1: Exogenous Confounder – Assume at first, or hypothesize, that the Blood Pressure plays the same role as Gender in the previous Causal Map #6. Blood Pressure is not affected by either the drug or the disease, but does affect both drug taking behavior and recovery outcomes. If this assumption is correct then Blood Pressure is a confounder, and the correct analysis is obtained by conditioning on this variable. The drug is harmful for both LBP and HBP, and should not be given to anybody. Different questions arise here: Why does the control group consist mainly of HBP patients, while the treatment group consists mainly of LBP patients? Perhaps HBP patients have some sort of adverse reaction to the drug, which leads them to stop taking the drug. Also: Why do HBP patients have much higher recovery rates, with and without the drug? Perhaps High blood pressure helps in recovering from the disease. Both of these hypotheses have real implications, and one could explore and experiment further to see if they are valid. This is a KEY aspect of causal chains. Causality itself is not observable, but it always has implications which can be explored and confirmed or rejected. Thus, indirect evidence is often available for hypothesized causal effects.

Figure 4.10 Causal Map # 7 A



When BP is an exogenous confounder, then conditioning on it gives the correct analysis. However, when it is a mediator, conditioning will give a wrong analysis, as we now demonstrate.

Exercise 8A: The standard way to handle confounders is via RCT – randomized controlled trial. Show that if we assign patients at random to treatment or control group, then both groups will

contain approximately the same proportion of LBP and HBP patients. Thus the effect of confounding will be eliminated. Note that this will be true even if the experimenter is UNAWARE that BP is a confounding factor. That is the beauty of randomization, that it eliminates even hidden confounding factors.

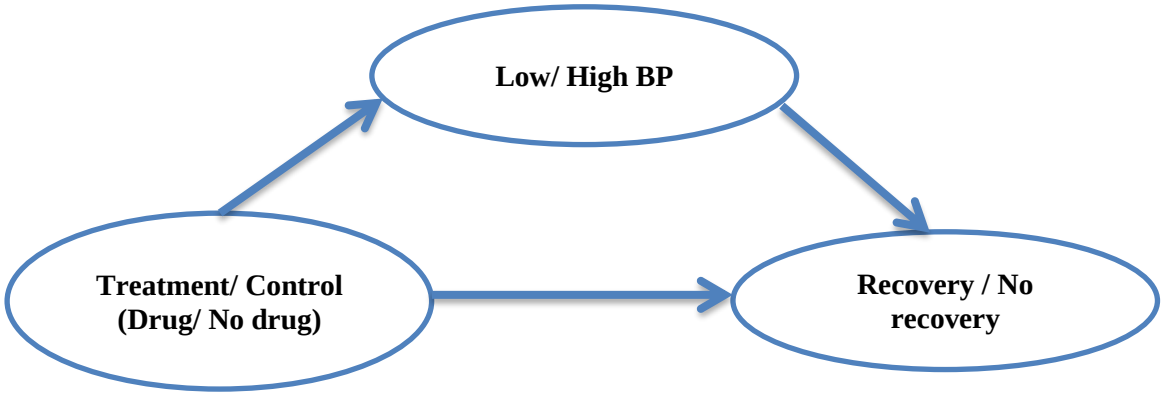
Exercise 8B: Suppose we do an RCT and find that there is a heavy concentration of High BP in the Treatment Group, but not in the control group. Explain why we can reject the hypothesis that BP is an exogenous confounder.

4.2. Blood Pressure as Mediator

Radically different analysis and interpretation of the data arise if Blood Pressure acts as a mediator, rather than an exogenous confounder. It is to illustrate this possibility that we switched from Gender to Blood Pressure, because Gender cannot be a mediator.

Case 2: BP is a mediator – Gender cannot be affected by Drugs, but Blood Pressure can be. There is a possibility that Blood Pressure acts as a mediator – a channel through which the drug works. The Low and High Blood Pressure may be one of the effects of the drug itself. If this is so, then the causal map looks like the following:

Figure 4.11 Causal Map # 7 B



Now BP is not a confounder because it is not exogenous. It is affected by the drug. This causal map suggests the following interpretation. High Blood Pressure is very helpful in recovering from the disease. The drug itself is toxic, and lowers recovery rates of all patients by about 20%. However, the drug causes an increase in blood pressure, which is very helpful in recovering from the disease. As a result, when we take the sum of both effects, we find that the drug is actually helpful. With this causal map, the drug is beneficial because of its indirect effect of raising blood pressure, which helps in recovery, even though its direct effect is harmful.

As we have noted earlier, all causal maps are actually guesses about hidden and unobservable structures of external reality. The data does not provide direct evidence for or against them, and all causal hypotheses are tentative, and remain tentative forever. As we demonstrated in the Berkeley admissions case, another hidden factor (like SAT) may upset all of our conclusions. However, our causal hypotheses are **productive**. That is they lead to additional hypotheses which can be tested. If Causal Map #7 is actually valid, we can go and look for mechanisms whereby high blood pressure helps in the recovery process from the disease. Finding this mechanism would support our causal hypothesis and may lead to discovery of other ways to help recovery which do not raise the BP. Similarly, we could look for the mechanism by which

the drug harms recovery. Search may lead to discovery of modifications of the drug which raise the BP without other harmful side effects. Even though the causal hypotheses are about unobservables, they have many observable implications which can be explored and tested to provide indirect evidence for and against the hypothesis.

Exercise 9A: Show that when BP is a mediator, conditioning on BP to calculate the effects of drug on recovery leads to the wrong results regarding the effect of the drug. In this situation, the overall effect in both groups gives the correct result regarding recovery rates, while the recovery rates in the separate groups of LBP and HBP give misleading results about the effect of the drug. Explain intuitively why this is so.

Exercise 9B: Given this data set and this causal map, we could try to predict recovery rates for people with HBP in the general population, who are not part of the experiment. What would these rates be? If we could find a drug which raised blood pressure without having any other side-effects, what would the recovery rates be for a group treated with this new drug which raises blood pressure from normal LBP to High HBP with 90% probability?

4.3. Simple Analysis for Simple Chains

Better understanding of the complex causal chains required for Simpson's Paradox arises from considering a simple chain of causation, which we illustrate below. In simple chains, one factor acts through the second only, and there are no multiple causal pathways. Understanding the simple case serves as a basis for understanding the more complex cases.

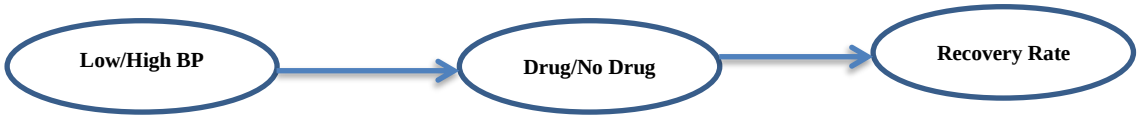
Case 3: BP as mediator – Acting on the hypothesis suggested by causal map #7, a detailed analysis of effects of the drug on human body is carried out. It is discovered that there are two compounds in the drug, one of which is toxic, while the other acts to increase blood pressure. A modified drug is created by eliminating the toxic compound. In this case, the drug has no direct effect on recovery, and ONLY acts to increase the Blood Pressure. However High Blood Pressure does have a strong positive impact on recovery. The Causal Map now becomes the following:

Figure 4.12 Causal Map # 8

Using the same numbers as before, suppose that in the LBP group, the disease is deadly with only 40% recovery rate, but in HBP group recovery rate is very high at 80%. Suppose in normal population 90% people are in LBP group and 10% in HBP group. Then the overall recovery rate in normal population without drug will be 44% = 90% *40% + 10%*80% = 36% + 8%. The new modified drug given to patients changes the BP profile so that 90% of patients now have High Blood Pressure, and only 10% remain at Low Blood Pressure. Even though the drug has no direct effect on recovery, this action of raising Blood Pressure helps recovery because of the helpful effects of HBP on recovery. Under our hypotheses, the new recovery rate will be 76% = 90% *80% + 10%*40% = 72% + 4%. By eliminating the toxic effect of the drug, the recovery rates in the treatment group will now soar to 76%.

Case 4: Drug as Mediator – The causal sequence in map #8 could be reversed. It may be that Blood Pressure governs whether or not the drug is taken, and has no direct effect on recovery.

Figure 4.13 Causal Map # 9



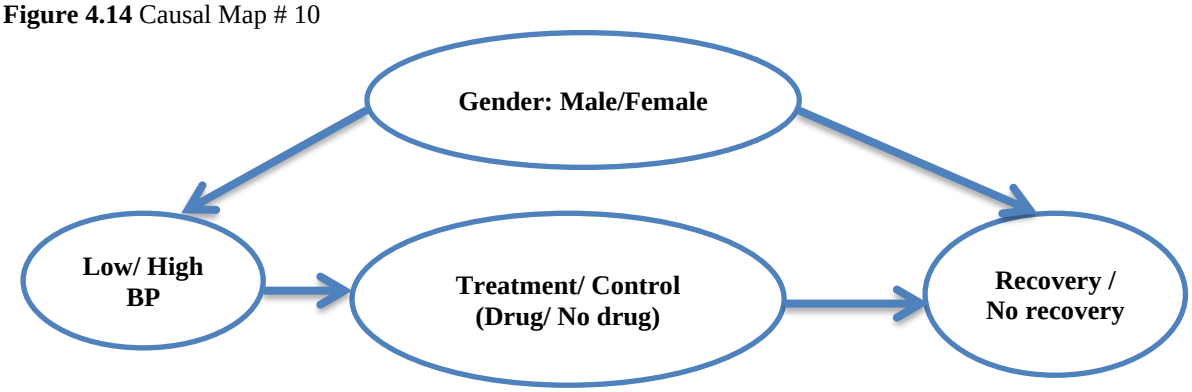
Suppose the drug is extremely helpful. Recovery rate in treatment groups (with the drug) is 80%, while recovery rate in the control groups (without the drug) is only 40%. Suppose the drug sometimes causes an adverse reaction which causes mild discomfort, which discourages people from continuing to take the drug. The adverse reaction occurs with very low probability in High Blood pressure patients, and with very high probability in low Blood Pressure patients. Depending on how the data is framed and presented, a statistician could easily come to the conclusion that High Blood Pressure leads to recovery, and Low Blood Pressure does not. Furthermore, if LBP is a large proportion of the population, the data could suggest that the drug only has a small positive effect, because most LBP patients abandon the drug regime.

Exercise 10: Suppose the situation is as described in the previous paragraph. The statistician creates a randomized controlled experiment dividing the population (which has 90% LBP and 10% HBP) into treatment and control groups at random. The treatment group is told to take the drug, while the control group does not take the drug. The drug tastes very bitter to 90% of LBP patients and they stop taking it after a while, but do not inform experimenter about their non-compliance. The drug tastes bitter to only 10% of HBP patients, who stop taking the drug. The experimenter does not know about the non-compliance. Calculate what his data will show about the effect of the drug, given that he assumes that everyone in the treatment group actually took the drug, and no one in the control group did.

4.4. Adding Confounder to Simple Chains

Simpson's Paradox is not possible in simple chains like the ones discussed above, because we need two pathways, one negative and one positive, for a reversal of effects. However, a confounder added to the above Causal Map # 9 can lead to a Simpson's Paradox:

Figure 4.14 Causal Map # 10



In this map, we do not actually have two channels by which BP affects recovery. However, if Gender affects BP, the two could be highly correlated, so that the data would have the appearance of showing two channels, when in fact the hidden variable was doing the work required for the second channel. This would lead to highly misleading interpretations of the causal mechanism relating the observed variables. In this particular case, we know, from our background knowledge of real-world mechanisms, that gender may affect BP, but the reverse

direction is impossible. The causal chain cannot run from BP to Gender. For other confounding variables this may or may not be true.

Exercise 11A: Suppose that BP has no direct effect on recovery. However, 90% of HBP patients continue to take the drug, while 10% stop taking drug because it tastes bitter. Among LBP patients, 90% of patients stop taking the drug because it tastes bitter to them, and only 10% continue to take the drug. In the general population, about 10% of the people are HBP and 90% are LBP. Recovery rates without drug are 40% (for all patients, LBP and HBP). Recovery rates for those who continue to take the drug are 80% for all patients LBP and HBP. Create a table which shows what the statistician will see if he studies a treatment group and a control group taken from the general population, but does not know about the fact that some people will stop taking the drug within the treatment group. What will the statistician conclude about the effectiveness of the drug on LBP and HBP patients?

Exercise 11B: Now suppose Gender is related to BP as follows. Among females, HBP is rare and only 1% of females suffer from HBP. Among males, HBP is far more common and about 20% of males suffer from HBP. In the same situation as that of the previous exercise, create a table for recovery rates by Gender in the treatment and control group. Ignore classification by BP, only classify by gender. What will the statistician looking at this table conclude about the effect of gender on recovery rates from the disease under treatment and without treatment?

4.5. Forks Lead to Correlation without Causation

There are a few more simple causal relationships which commonly occur among. One of them is called the "fork". Two possible fork relationships can be depicted as follows:

Figure 4.15 Causal Map # 11 A

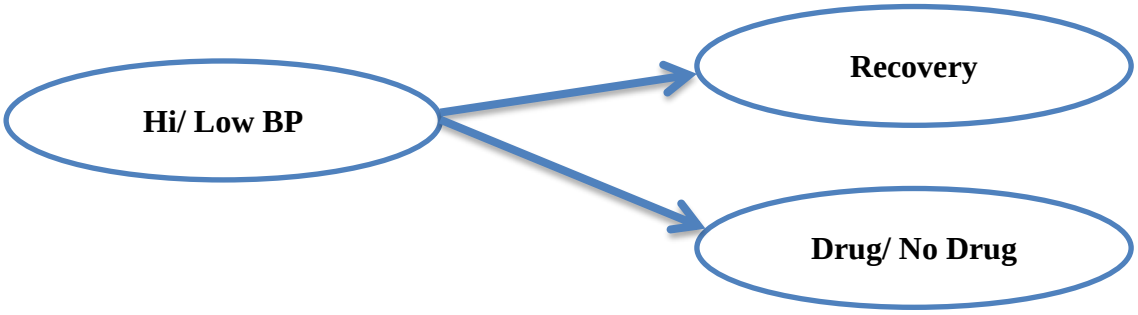
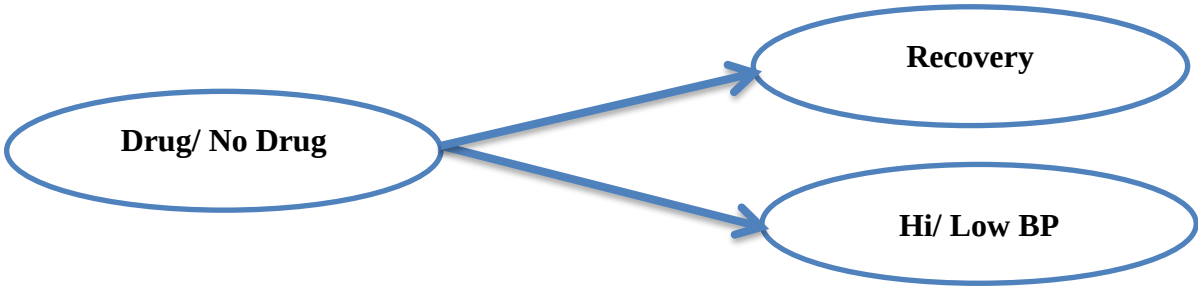


Figure 4.16 Causal Map # 11 B



In the first causal map 11A, the Blood Pressure is the only factor which matters for recovery. However, the Blood Pressure also affects whether or not the drug is taken; as before, the drug may have a mild adverse effect on High or Low BP patients, creating differential patterns of drug taking. Statisticians analyzing the relationship between drug taking and recovery rates would come up with highly misleading conclusions if they were not aware of these causal relationships. Similarly, in causal map 11B, the drug affects both recovery and Blood Pressure. With this type of causal chain, almost any kind of relationship between Blood Pressure and Recovery could be observed in the data, when there is none in reality.

Forks create correlations which are misleading. In the first map above, strong correlation between drug taking and recovery would lead to the impression that drug taking has a strong effect on recovery. In fact, there is no effect. Similarly, in the second map, recovery would be highly correlated with recovery and yet neither factor causes the other. A Granger type causality test between the two factors would lead to highly misleading results.

Exercise 12: Suppose that, as before, 90% of HBP patients in treatment group continue to take the drug, while 10% stop taking the drug. Among LBP patients, 90% stop taking the drug while 10% continue to take it. Suppose that the drug has no effect on recovery, but HBP is a strong positive factor. Among LBP patients, recovery rate is 40% with or without the drug. Among HBP patients, recovery rate is 80% with or without the drug. Suppose both control group and treatment group are chosen at random from the general population which has 90% LBP and 10% HBP. There is no effect of gender.

a) Calculate the overall recovery rates in treatment and control groups.

b) Calculate recovery rates among HBP and LBP separately in both treatment and control group.

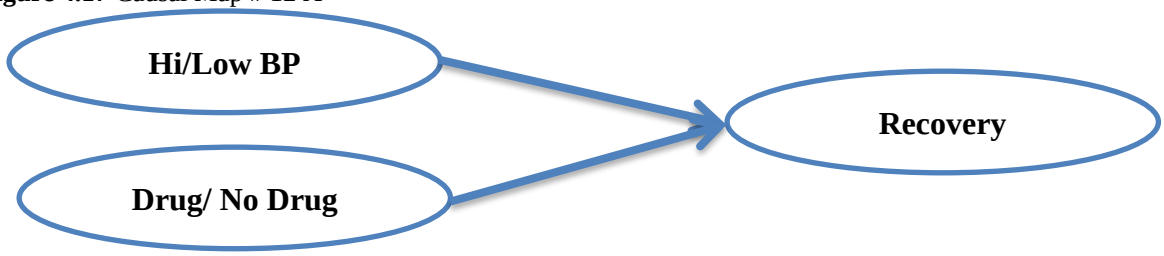
c) What are the conclusions that statisticians who analyze the data might come to, given that they do not know the causal mechanism is a fork?

d) What kind of an experiment would reveal that the causal mechanism is a fork?

4.6. Do NOT Condition on Colliders

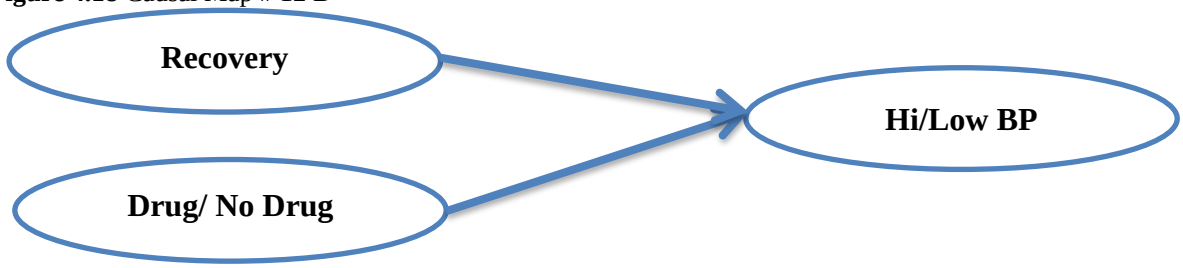
A final example of a simple causal relationship among the three variables is called the "collider". This type of causal relation is on the other side of the causal map from the confounder. Whereas a confounder affects the variables under study, a collider is affected by the variables under study. The significance of colliders lies in the fact that they require different treatment from confounders. While one must condition on confounders for correct analysis, one must NOT condition on colliders for correct analysis. In the example under study, there are two possible colliders:

Figure 4.17 Causal Map # 12 A



In the map 12A, Recovery is a collider, which is affected by both BP and Drugs. A second possibility for a collider map is the blood pressure, as picture in causal map 12B below:

Figure 4.18 Causal Map # 12 B



Note that the third possible collider is ruled out because Recovery cannot be a cause of taking the Drug or not. Causal chains depend strongly on our knowledge of external reality – we could not rule this out if we were considering variables X, Y, Z without knowledge of what the variables represent.

Causal Path 12A: The first collider is a case we have already encountered. In this the Hi/Lo BP acts as a confounder in assessing the relationship between Drug Treatment and Recovery. If the Blood Pressure is ignored, we can get entirely wrong assessments about this relationship. Similarly, someone studying the relationships between Blood Pressure and Recovery without considering causal effects of the Drug will arrive at the wrong conclusions. In the context of regression models, this point has been emphasized in Zaman (2017) "Choosing the Right Regressors" – omitting a significant regressor will lead to misleading regression results.

Causal Path 12B: This is an interesting case where recovery from the disease occurs naturally, without drugs, but leaves some effect in terms of higher or lower Blood Pressure. The drug treatment has no effect on the disease, but it also affects Blood Pressure. To consider an extreme case, suppose all patients who recover end up with High Blood Pressure, and similarly all patients who take the drug end up with High Blood Pressure. Analyzing the patients with HBP, we will see that all patients took the drug and recovered. Analyzing patients with LBP we will see the no one took the drug and no one recovered. So, we will conclude that the drug is highly effective, even though it has absolutely no effect on recovery. Thus, conditioning on the collider leads to a highly misleading analysis.

Exercise 13: Consider the causal path 12B and suppose that natural recovery rate from disease is 60%, with or without the drug. In the general population, LBP is 90% and HBP is 10%. However, the disease causes elevated blood pressure which persists even after recovery from the disease. Among those who recover from the disease HBP is 80% while only 20% have LBP. The drug has no effect on the disease, but it also has the same effect on the blood pressure. Taking the drug leads to HBP in 80% of the patients and LBP in only 20%. Show that a statistician who analyzed recovery rates in a randomized control group and treatment group without taking BP into consideration will come to the right result. However, if he thinks that BP is a confounder, and splits the treatment and control group by BP, he will get wrong results about the impact of the drug.

5. CONCLUSIONS

From the last example we see that when analyzing the relationship between Drugs and Recovery from disease, the Blood Pressure can play three different roles. It can be a confounder, which affects both drug-taking and recovery, without being affected by either. In this case, conditioning on BP leads to the correct conclusions about the effects of drugs on recovery. The BP can also be a mediator, or a channel by which the drug affects recovery. In this case, BP is endogenous and conditioning on it leads to the wrong results. The right analysis involves taking into account both the direct effects of the drug, and the indirect effects via the BP channel. The third possibility is that BP is affected by both Drugs and Recovery, but does not affect either of the two. In this case, it is a collider, and should be ignored in the analysis. If we take it into account, either as a confounder, or as a mediator, we will get wrong and misleading results. Which of the three causal path diagram obtains is not possible to detect directly from the data. This means that data analysis cannot be confined to the study of numbers, without considering the underlying and hidden causal structures of external reality.

This discussion carries over to the regression context. Typical econometric studies involving three variables *X, Y, Z* would only run one of the variables on other two. As we have seen, the number of causal path relationships possible among the three is much larger. Even for standard regression analysis, we first need to determine exogeneity among these variables. Now this is not possible without understanding the real-world causal structures which generate the data. Econometrics does not provide us with any formal language to represent causal relationships. Causality can only be discussed qualitatively and informally, and cannot be represented mathematically in the equations we write. This is why econometrics textbooks (and students) are massively confused about this concept.

It is evident from the above given discussion that causal relationships in real life can actually be very complex. The hard work involved in the process of searching for causes has been described by Freedman (1991) in several real-world examples. Because econometricians are not taught to think about causes, most of the regression relationships we write down are spurious. Without additional information, unobservable real-life causal structures cannot be understood and without understanding the causal structures, exogeneity of the variables cannot be determined. The deficiency comes from a ‘positivist’ methodology which consciously limits analysis only to observables. By excluding non-observable causal patterns from consideration, econometric methodology prevents us from learning about causal relations. To make progress, it is essential to discuss causal structures, which lie beyond the numbers, and depend on our qualitative knowledge of external reality.

REFERENCES

Bickel, P.J., E. A. Hammel and J.W. O'connell, (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*. 187 (4175), 398-404.

Freedman, D. A. (1991) Statistical Models and Shoe Leather. *Sociological Methodology*, 21, 291-313.

Hoover, K. D. (2004). Lost Causes. *Journal of the History of Economic Thought*, 26 (2), 149-164.

Pearl, J. and D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Pearl, J., M. Glymour and N. P. Jewell (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Zaman, A. (2010). Causal Relations via Econometrics. *International Econometrics Review*, 2 (1), 36-56.

Zaman, A. (2012). Methodological Mistakes and Econometric Consequences. *International Econometric Review*, 4 (2),.99-122.

Zaman, A. (2017). Choosing the Right Regressors. Lessons in Econometric Methodology: The Axiom of Correct Specification. *International Econometrics Review*. 9 (2), 50-68.

Zaman, A. (2018). European Transition to Secular Thought: Lessons for Muslims. https://sites.google.com/site/azamanpublications/ie/islamic/european (accessed April 10, 2020).