

APPENDIX A

COVER SHEET TO BE APPENDED TO ALL THESES

By law, copyright in your thesis belongs to you, the author, including all rights of publication and reproduction. Only the copyright holder may determine what others may do with the thesis.

STEP 1. What should the library do with your thesis? (Choose one)



☒ Share my thesis

You authorize the library to share your thesis with others according to the Creative Commons license selected in step 2. (See Appendix B for an explanation of the three options.)



☐ Place an optional embargo on my thesis

The library should not allow anybody except college officials to see its copy of my thesis until January 1st of the year _____. (e.g., 1 year=2018, 5 years=2022, 100 years=2117)

Under this option, you ask the library to prevent anyone else from accessing its copy of your thesis. At the end of the lockdown period, the library will distribute its copy of your thesis according to the license you choose below. (See Appendix B for an explanation of the three options.)

STEP 2. Select a license for your thesis. (Choose one)

☐ CC BY-NC-ND

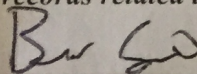
☐ CC BY-NC-SA

☒ CC BY

The type of Creative Commons license you choose determines what others may and may not do with your thesis. (See Appendix B for an explanation of the three options.)

STEP 3. Sign.

Please note that, by choosing to make your thesis available (whether now or at a future date), you are waiving any protections of the Family Educational Rights and Privacy Act (FERPA) that may apply with respect to your thesis as of the date specified. FERPA generally restricts disclosure by the college of records related to your education.



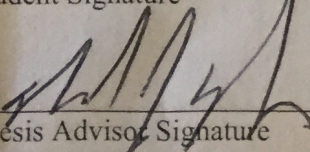
Student Signature

Brendan Seto

Student Name

5/1/18

Date



Thesis Advisor Signature

Nick Horton

Thesis Advisor Name

5/1/18

Date

Appendix B: Creative Commons Licenses

CC BY-NC-ND

Attribution-NonCommercial-NoDerivs



This license is the most restrictive, only allowing others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially.

- License Deed: <http://creativecommons.org/licenses/by-nc-nd/3.0/>
 - Legal Code: <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>
-

CC BY-NC-SA

Attribution-NonCommercial-ShareAlike



This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.

- License Deed: <http://creativecommons.org/licenses/by-nc-sa/3.0/>
 - Legal Code: <http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode>
-

CC BY

Attribution



This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

- License Deed: <http://creativecommons.org/licenses/by/3.0/>
 - Legal Code: <http://creativecommons.org/licenses/by/3.0/legalcode>
-

More information about Creative Commons licenses is here:

<http://creativecommons.org/licenses/>

Causal Inference

Brendan Seto

May 1st, 2018

Submitted to the
Department of Mathematics and Statistics
of Amherst College
in partial fulfillment of the requirements
for the degree of
Bachelor of Arts with honors

Faculty Advisor: Dr. Nicholas Horton

Copyright © 2018 Brendan Seto

Corrections

When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected.

Various places in the thesis. Approximately 7 typographical errors were corrected in the thesis proper, primarily changing principal to principle.

Title Page Changed formatting to conform with other Statistics Thesis writers.

- p. 9. Added more detailed figure caption specifying that coffee is the treatment and cancer is the outcome.
- p. 9. 1. 6. Replaced “by Sewall Wright’s book” with “by Sewall Wright in his 1921 book”
- p. 10. Added more detailed figure caption specifying that coffee is the treatment and cancer is the outcome.
- p. 11. Added more detailed figure caption specifying that coffee is the treatment and cancer is the outcome.
- p. 14. Clarified Association in Table 1.1 between coffee and cancer.
- p. 14. Added more detailed figure caption specifying that coffee is the treatment and cancer is the outcome.
- p. 14. 1. 14. Added in “dichotomous” to clarify that the treatment has only two levels.
- p. 18. Referred readers to appendix to see different definitions of confounding in footnote.
- p. 21. 1. 12. Added additional citation.
- p. 24. 1. 19. Deleted “very”
- p. 29. Replaced references to “sugary drinks” with references to number of cigarettes.
- p. 34. Clarified that simulation based on true differences, but exaggerated slightly for clarity of point.
- p. 34. Added footnote explaining previous choice of using binary smoking variable as opposed to the more informative number of cigarettes variable.

- p. 44. Added line “This method of propensity score adjustment is a common, effective and user-friendly way to identify a causal effect, even in the presence of confounders in observational data.”

Acknowledgements

Thank you to Nick for guiding me through this thesis process. Without his mentoring this project would be impossibly hindered by procrastination and a general sense of uncertainty. With him, it was instructive and a lot of fun.

Special thanks also to Sarah Anoke, who took time out of her busy post-doc life to answer my questions and comment on my draft.

Finally, thank you to my friends, family and classmates who have put up with seeing me even less than they normally would have during hibernation season.

Table of Contents

Introduction	1
Chapter 1: What is Causality?	3
1.1 Historical Perspective	4
1.1.1 In Philosophy	4
1.1.2 In Statistics	5
1.1.3 In Biomedical Research and Epidemiology	6
1.2 Formal Perspectives	8
1.2.1 Counterfactuals: Basis of Causal Inference	8
1.2.2 Visualizing Causal Relationships with Directed Acyclic Graphs	9
1.2.3 D-separation	12
Chapter 2: Estimating Causal Effects	17
2.1 Using RCT Data	17
2.1.1 Confounders: A Lurking Danger	18
2.1.2 Effect Modification	19
2.1.3 Selection Bias	20
2.1.4 Sampling Bias	22
2.2 Using Observational Data	23
2.2.1 Identification Assumptions	23
2.2.2 Validity of Resulting Causal Conclusions	25
2.3 Methods to Determine Causal Effects	26
2.3.1 Stratification	27
2.3.2 Confounder Matching	27
2.3.3 Weighting	28
2.3.4 Regression Control	29
2.3.5 Propensity Scores	30
2.3.6 Instrumental Variables	31
Chapter 3: Simulation Study and Applications	33
3.1 Simulation Study	33
3.1.1 Creating A Hypothetical Population	33
3.1.2 The Search for a Confounder	33
3.2 Applications	38
3.2.1 Clearly Wrong: Bone Marrow Transplant	38

3.2.2	Causality is Difficult: Hormone Replacement	40
3.2.3	Real Data Example	41
3.2.4	Conclusion	44
Chapter 4:	Conclusion	47
Appendix	49
4.1	Definitions From Various Texts	49
4.2	Code Used to Create Simulation Study Population	49
References	53

List of Tables

1.1	Testing Potential DAGs. Suppose there is no direct causal effect between coffee and cancer (if there was, there would always be an observed association). There will still be an observed association in DAG C through F. However, only D and F depict a true causal effect. By only controlling for smoking, they cannot be distinguished from their non-causal counterparts.	16
2.1	Sir Bradford Hill's criteria to determine causality (1965).	26
3.1	Population characteristics of the simulation study. Values based on true phenomena, exaggerated slightly to allow for more obvious conclusions.	34
3.2	Sample of true population data in simulation study. One of the primary advantages of a simulation study is that both counterfactuals are known. Thus, we can calculate a true effect for each individual and measure the accuracy at each step in our analysis.	34
3.3	Sample of observable population data in simulation study. Only one counterfactual is visible now. Thus, to get an estimate of the causal effect, we need to use the observed counterfactual of the control group to approximate the unobservable counterfactual of the treatment group.	35
3.4	Results from two sample t-test comparing observed cancer risk in coffee drinkers to non-drinkers. Stratifying by sex reveals the presence of an effect modifier, but coffee still appears to cause cancer.	35
3.5	Initial models to describe cancer risk score. Notice that coffee drinking is initially significant, but is not after smoking status is added to the model (marked in red). This shows that smoking is a mediator or common cause in the relationship between coffee consumption and cancer.	37
3.6	Models after accounting for dose effect. Cancer risk was initially generated by a linear combination of the patient's baseline cancer risk and the number of cigarettes they smoke per week. Thus, although sex was significant when smoking was measured as a dichotomous indicator, it was not when the more specific measurement was used. Moreover, once the regression model was left with only the number of cigarettes variable, the generating model was recovered, with the intercept serving as the average baseline cancer risk score in the population.	39

3.7	Hypotheses from observational tests after further study in RCT. Some have proved true, others have not. Example taken from Dr. Michael Lauer's presentation at the International Society for Pharmacoeconomics and Outcomes Research.	40
3.8	Initial models from real data: hour only and hour with patient characteristics. Once patient characteristics were added in, time of day lost significance in both the mortality and complication rate models. . . .	43
3.9	Propensity score as coefficient adjustments. Models including propensity scores perform similarly to those controlling for each covariate separately.	45
4.1	Definitions of confounding and discussion of sampling or epidemiological selection bias from various texts.	52

List of Figures

1.1	A review of PubMed, an online repository of medical publications. There has been an exponential growth in studies based on observational data. Randomized trials, on the other hand, display a more linear growth. There are currently far more studies based on observational data than randomized trials	7
1.2	Simple Directed Acyclical Graph (DAG). The arrow indicates that the variable at its origin, X, is a direct cause of the variable at its terminating end, Y.	9
1.3	Mediator example. A mediator is an intermediary even in a causal chain linking the outcome and treatment. Mediators are active in their natural state. Therefore there is an observable association between the outcome and treatment, unless the mediator is conditioned on. In this example, Coffee (X) is a cause of Cancer, but only through the mediators Cafe and Smoke.	10
1.4	Common cause example. A common cause is a cause of both treatment and outcome. It is active in its natural state. Therefore there is a spurious association between the treatment (Coffee) and outcome (Cancer) in this graph, unless the common cause (Smoking) is conditioned on.	11
1.5	Collider example. A collider is a variable that is an effect of both the treatment and outcome. It is inactive in its natural state, meaning there is no association between the treatment (Coffee) and outcome (Cancer, through smoking) in this graph. Conditioning on a collider (Yellow Teeth), however, will activate it and induce a spurious association.	12
1.6	D-separation example containing a common cause (red path), mediator (green path) and common effect (blue path)	13
1.7	DAGs in use. Dotted gray arrows with question marks indicate the (unknown) relationship of study. In this DAG, coffee is the treatment while cancer is the outcome.	14
1.8	DAGs illustrating the potential relationships between coffee, cancer and smoking. L represents any other lurking variables that may lie on the causal chain between the treatment (coffee) and effect (cancer).	15
2.1	Confounder example. Here we see that L is a confounder of X and Y. While L remains uncontrolled, it is impossible to make a determination on the relationship between X and Y, if one even exists.	18

2.2	Randomization works by breaking the connection between assignment of treatment and any external covariate. Thus, it eliminates potential confounders and minimizes the risk of bias.	19
2.3	Selection bias example. Selection bias arises from improperly conditioning on a common effect such that a spurious association is induced. Examples include loss to follow up and noncompliance	21
2.4	Instrumental variable example. An instrumental variable allows researchers to avoid confounding by using an external covariate to predict an untainted expectation of the treatment. While generally not used by epidemiologists and statisticians, it is a favored method of economists and other social scientists.	31

Abstract

Causal inference is an essential discipline in statistics. By finding a causal relationship between two or more variables, investigators can determine how and when to intervene to force a certain outcome to occur. In statistics, medicine, and life in general, this idea is the underlying mechanism that fuels many decisions. Yet some people neglect to consider causal inference from studies with non-randomized assignment of treatment. While randomized controlled trials are the gold standard for determining causality, they are just the simplest method of doing so. This thesis aims to build a basic understanding of the foundational concepts of causal inference by distilling the literature into digestible pieces. By defining core terms, introducing common families of methods, walking through a simulation study, and exploring real life applications, students can gain a preliminary understanding of this broad topic. These ideas can then be developed in whatever field of study students decide to pursue so that they are prepared to analyze data from a complex and multivariate world.

Introduction

We live in a wild, stochastic, and evolving world. Many of humanity's greatest achievements have been built upon the principle of adapting this environment to our advantage. But nature's mechanisms are a black box, taking in an input, X , and producing an observed output, Y . To unravel these mysteries, we developed scientific inquiry, which allows us to reveal Mother Nature's inner workings. With this knowledge, we can devise treatments to promote the occurrence of certain beneficial outcomes which better our lives. This process of unveiling the mechanisms of the world to understand how and when to intervene is known as causal inference.

Historically, there has been some debate among statisticians on how best to approach causal inference (Holland, 1986). While statisticians generally agree that the stochastic nature of causal problems requires an understanding of probability and statistics, actually estimating unbiased causal effects is difficult. Randomized control trials (RCT) are widely accepted tools to demonstrate causal connections, but even they are limited in their use and scope. Sometimes the needed data are too time consuming and costly or simply impossible and unethical to collect. For example, imagine a RCT to determine the effect of smoking on cancer incidence. This study cannot be blinded as patients will know which group they belong to. Also, most researchers would feel uneasy assigning a person to smoke cigarettes when the likelihood of putting participants at an increased risk of cancer is high. Even without these limitations, it is very difficult to continue RCTs over decades to assess long-term outcomes. Thus, while RCTs may be the best way to estimate causal relationships, their use is limited.

Moreover, we live in a world made up of primarily “found” or observational data. Due to new technologies, such as large sensor networks or electronic medical records, large observational data sets are becoming increasingly available. Drawing causal inferences from these kinds of data is fraught with difficulty, often relying on a researcher's willingness to accept certain untestable assumptions. Although it is inevitable that people will interact with these data or depend upon conclusions drawn from such sources, many still preach the mantra “No causation without manipulation.” Such a narrow perspective may leave prospective researchers ill-equipped to contribute in the era of big data and may in fact lead to an over-reliance and blind acceptance of results that lack validity or generalizability, simply because there was an element of random treatment assignment. Readers must have a fundamental understanding of the essence of causal inference to evaluate scientific claims.

The past two decades have also brought about huge advancements in machine

learning and other computational big data techniques. Advances in these fields allow us to generate predictions from the “black box” without needing to worry about causality. Many industries such as finance and technology have used these techniques to great effect and are now seeking to hire more statisticians.

Statistical education, therefore, has begun to place a higher premium on these skills, with “big data” and “data science” becoming the hottest keywords on many students’ resumes. Yet big data techniques on their own often produce associational, not causal, relationships. Taken in context with the general lack of focus on analytic approaches to observational data, statistical education is simultaneously training students in skills to identify association while driving into their minds the idea that only RCTs can be trusted. In truth, this is not an insurmountable divide. Much of the knowledge necessary for causal inference is qualitative in nature. Students often simply need to better understand how and when to apply the skills learned in introductory courses. By putting their data in context and developing an understanding of causal structures, students at a variety of levels can begin interpreting their results to reach causal conclusions. If students do not have this ability, they are fundamentally unprepared for real-world analysis, no matter how technically gifted they are. Causal inference needs to be a fundamental aspect of statistical education.

Students need to be prepared to live in and understand a complex, multivariate world of mostly observational data. We believe it is imperative that they understand the basic concepts of causality in order to do so. This thesis is intended to foster an understanding of causal inference and its most prominent methods. We start with a basic overview of causal inference designed to familiarize a reader who has some background in statistics with the fundamentals of causal inference. We will then explore experimental design to gain an understanding of how causal relationships are estimated so that we can replicate this process in observational data analysis. Following that will be a historical overview of approaches to causal inference as well as an overview of common methods. Finally, we will conclude with a simulation study and discussion of several real world applications.

Chapter 1

What is Causality?

Causal inference is based on a deceptively simple idea. Given a treatment, X , and outcome, Y , we say that X is a *cause* of Y if changing the state of X will result in a corresponding change in Y (Rubin, 1974). Anytime we want to make something happen, causal inference is needed to figure out how to cause the change we need. It turns out that many questions that arise in medicine and the social sciences are causal in nature. We do not just care about what is wrong, we need to figure out how to fix it. Note that there is a difference between this concept and association. *Causation* tells us where to intervene to achieve a certain outcome. *Association* refers to two variables, X and Y , having a linear relationship with each other. In other words, in a given population, knowing the value of X provides some information about the value of Y (Rubin, 1974). Yet this does not tell us anything about the relationship between the two variables. While it is true causation implies association, the reverse is not necessarily true. Association can also be produced by Y causing X , a common variable causing both, further knowledge about a common effect, or random chance. Section 1.2.2 will explore these relationships in a more formal context.

The difference between causation and association may be most evident in an example. Say a patient walks into a doctor's office complaining of chest pain. The doctor may correctly recognize the patient exhibiting early signs of a blocked coronary artery, and deduce through association that the patient is about to have a heart attack. Here we see both association and causation. Patients with chest pain are significantly more likely to experience heart attacks, thus it is a useful diagnostic measure. But, giving a patient Tylenol to treat the chest pain will not necessarily reduce the risk of a heart attack. Thus, intervening on chest pain will not treat their heart attack and there is no causal relationship. Contrast this with a blocked coronary artery. Patients with blocked coronary arteries are more likely to experience a heart attack. Thus, we have an association. But, if we were to intervene and clear out their coronary arteries, the patient's risk drops dramatically. This shows a causal relationship.

Causal inference is fundamentally about determining where and how to intervene in order to achieve a certain outcome. Yet, like much of statistics, actually calculating and showing this aim proves much more difficult than it appears. This section will detail several basic concepts that are integral to causal inference efforts and give readers a foundation from which they can begin to tackle these problems primarily

following the notation and approaches of Scheines (2017) and Hernán and Robins (In progress).

1.1 Historical Perspective

The history of causal inference is long, multidisciplinary and full of spirited debate. In a way, the fundamental principle behind causal inference, figuring out what to change to yield a certain outcome, is the backbone of science and discovery as a whole. Babies learning to take their first steps are doing causal inference just as much as a researcher testing a new drug to cure HIV. A dog learning that sitting leads to a treat has done so as well, even if it cannot verbalize what it is doing.

Given the broad application of causal inference, this literature review will begin with a selective account of the development in causal inference in three distinct fields: philosophy, statistics and medicine. While these descriptions may be a slight tangent, they highlight the need for causal inference and where it came from.

1.1.1 In Philosophy

As with much of science, the first evidence of formalized causal inference stems from Aristotle's teachings in Plato's *Republic* (381 BC). Aristotle emphasized that we truly needed to understand why something happened, as opposed to just what had happened. Only then, he reasoned, would we be able to influence it.

The topic came into focus again much later, when philosophers began to question if we can ever truly know a cause (or anything for that matter). It seemed to them that there may be no "true" causes. Yet the philosopher Hume countered this idea by proposing a theory that states: even if it is not possible to empirically prove that an event causes another, it is enough to say that one event invariably follows another (Hume, 1740). In practice this is equivalent to our current definition, as changing the state of the first event is sufficient to affect the second.

Over two centuries later, Patrick Suppes (1970) accounted for the stochastic nature of the world brought about by indeterminism - the theory that there is always some degree of randomness in nature. If the exact same sequence of events occurs, even down to the subatomic level, there is no guarantee that the event that follows will always be the same. But, if we can never say for sure that intervening on X will affect Y , our present definition does not hold. This is where Suppes steps in. Even if we are not 100 percent sure that X will follow Y , Suppes relaxes the definition of causality to require only that the probability of an unpreferred state of Y becomes preferred under the changed value of X . In other words, changing X will most likely change Y . In adapting this definition, Suppes formally injects probability and statistics into causal inference (Suppes, 1970). The field of causal inference in statistics begins to bloom in the decade after Suppes' paper.

1.1.2 In Statistics

Causality has been a controversial aspect of statistics since the discipline's modern inception. Two of the early pioneers of the field, Francis Galton and Karl Pearson (1920), have been credited for formally defining correlation (X changes in proportion to Y). Yet they chose to take the most conservative standing in rejecting any notion of causation beyond correlation. This stance was so strong and unyielding that Judea Pearl (2009) would eventually refer to Pearson as “causality’s worst adversary”.

Several years after Pearson’s discoveries, Fisher described the beauty of randomized assignment of treatment that laid the groundwork for randomized trials and comparative inference (Fisher, 1920). Fisher’s work was well received, but at this point causality was still restricted to a small subset of studies generated under certain conditions.

It was not until the 1970s that the causality model was fully developed to consider observational data (Rubin, 1974). One of the early pioneers of the field was Donald Rubin, with some later works titling his framework for causality the “Rubin Model” (Holland, 1986; Little & Rubin, 2000; Reiter, 2000; Rubin, 1974). Once the basic principles were established, several others joined him to expand and refine the subject.

One of the influential voices in the field is Judea Pearl, a computer scientist who was well regarded for his prior work developing Bayesian networks. Among his notable contributions are establishing the mathematical methods for causal discovery, creating a back-door criterion to guarantee bias-free estimation and allowing for systematic removal of potential confounders (Russell, 2011).

James Robins is another scientist who must be discussed when talking about the history of causal inference. As the second-most published author in the field (behind only Rubin), he has made a variety of contributions that have shaped the way researchers approach causal problems (An & Ding, 2017). Some of his most significant works include the establishment of structural nested models and G-estimators, which improve upon old methods by allowing for time and past history dependent risk factors. Pearl and Robins have worked extensively with each other, in some ways sharing the same sphere of influence (Harvard Faculty Profiles, 2018).

More recently, statistician Mark Van der Laan has made a name for himself in modernizing the field, with several of his most cited papers focusing on adapting machine learning methods to answer causal questions. He is also a founding editor of the *Journal of Causal Inference* (UC Berkeley Faculty Profiles, 2018) and, in 2005, won the Committee of Presidents of Statistical Societies’ highest honor for his “success in bringing statistical rigor into many fields of the biomedical sciences” (Past Winners, 2018).

Richard Scheines is another notable causal scientist. Interestingly, he has served as the chair of the Department of Philosophy at Carnegie Mellon University’s Dietrich College since 2005 and is remarkable for his multidisciplinary approach (CMU About the Dean, 2017). After making many contributions early on in the development of causal inference, he has recently made a significant effort to make the discipline more accessible through free online projects. One such effort is the Tetrad Project, an application that puts many of his notable contributions into a user-friendly interface

and allows people without much computing background an opportunity to explore central concepts (Scheines, 2017). He also has an open source course on Causal and Statistical Reasoning with Peter Spirtes that may serve as a useful introduction for those interested in the field (Spirtes, 2017).

Other researchers have recently taken Scheines' lead and made a noticeable push to make causal inference more accessible. Miguel Hernán, a professor with dual appointments in the departments of Epidemiology and Biostatistics at the Harvard Chan School of Public Health, also has an open source online course, "Causal Diagrams: Draw Your Assumptions Before Your Conclusions", and is collaborating with Robins to write an accessible texts on the subject (Hernán & Robins, in progress). Another accessible and well-known text is "Counterfactuals and Causal Inference" by Morgan and Winship (2007), written and used by social scientists. These accessible materials have opened up a previously dense topic and exposed it to a wider audience. In many ways, they serve as inspiration for this thesis.

1.1.3 In Biomedical Research and Epidemiology

Medicine has been one of the primary beneficiaries of the development of causal inference and is a useful example of the field's evolution. Thus, we will introduce some history of causality in medical research and use health related examples to illustrate our points.

Medicine began as a master-apprentice tradition. Physicians learned and trained by following an established master doctor, who in turn learned from a master many years prior. Diagnoses and treatment decisions were based primarily on physicians' own experiences or those of their teacher. This style of learning persisted until the 1980s (Sackett, 1981). By this point, people began to see the superiority of experimentally proven results and realized that the vast amount of new knowledge exceeded the capability of any one person to master it. Thus medicine finished a long transition from craft to scientific discipline.

The advent of the internet allowed distant collaborations and easy sharing of results. This accelerated the dispersion of scientific evidence and allowed medical practitioners around the globe to benefit from the discovery of others. The status quo became evidence-based medicine, the use of scientific evidence to make treatment decisions (Guyatt, 1991).

Traditional practitioners of evidence-based medicine give greater credence to findings from randomized control trials. Yet the limited number of randomized trials cannot encompass all medical findings. In a stunningly satirical paper published in the *British Medical Journal*, Gordon Smith (2003) wrote a review article hoping to find a RCT to back up the traditional practice of deploying a parachute to prevent "major trauma related to gravitational challenges". His search came up with no results. The conclusion to the paper highlights the impracticality of only following results from randomized trials:

Conclusions As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evalua-

tion by using randomized controlled trials. Advocates of evidence-based medicine have criticized the adoption of interventions evaluated by using only observational data. We think that everyone might benefit if the most radical protagonists of evidence based medicine organized and participated in a double blind, randomized, placebo controlled, crossover trial of the parachute (Smith, 2003).

Researchers have now begun to figure out ways to integrate non-experimental data sources into scientific literature. In some ways, this change was inevitable, as experiments are costly and time consuming. The same cannot be said about observational studies, given the huge amount of available data (Figure 1.1).

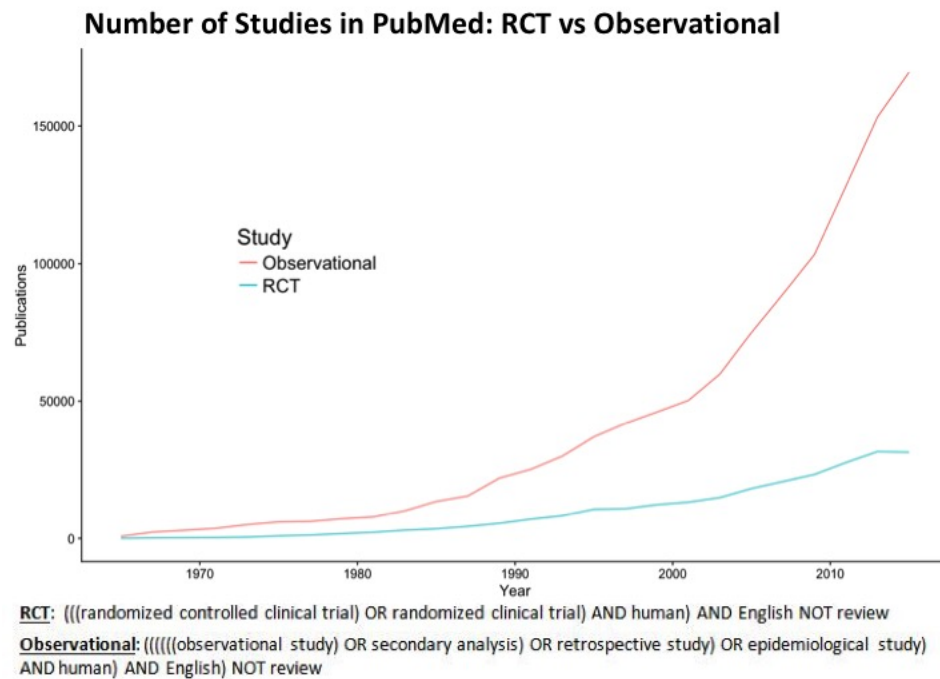


Figure 1.1: A review of PubMed, an online repository of medical publications. There has been an exponential growth in studies based on observational data. Randomized trials, on the other hand, display a more linear growth. There are currently far more studies based on observational data than randomized trials

Recent publications have also taken a notable shift towards educating the medical community about conclusions based on observational data. As recently as August 2017, the *New England Journal of Medicine* published a review article titled: “Evidence for Health Decision Making - Beyond Randomized, Controlled Trials”. In it, Thomas Frieden, a former director of the CDC, painstakingly details the advantages and limitations of data from a variety of sources (Frieden, 2017).

1.2 Formal Perspectives

1.2.1 Counterfactuals: Basis of Causal Inference

We have now discussed the general principle behind causality. Given a treatment, X , and outcome, Y , we say that X is a *cause* of Y if changing the state of X will result in a corresponding change in the expected value of Y . This definition leads naturally to the foundational method of causal inference: comparing outcomes in an individual that does or does not receive treatment. If the outcome changes, the treatment had a causal effect.

Suppose we want to determine the effect of a dichotomous treatment, X , on an outcome variable, Y . Denote the observed outcome under treatment as $Y^{X=1}$ and the outcome without treatment as $Y^{X=0}$. These values (Y^x) are known as *counterfactuals* or *potential outcomes*. To find a causal effect, we need to determine if $E(Y^{x=1}) \neq E(Y^{x=0})$. Put simply, we compare the expected value of the two counterfactuals. If their expected values are different, on average X will have a causal effect on Y (Rubin, 1974).

Although we use $E(Y|X)$ to estimate $E(Y^X)$, the two sets of data being averaged are different. This is a consequence of the fact that there is a distinction between the counterfactual population $Y^{X=x}$ and $Y|X = x$. The first is the population if every individual were to receive treatment; the later is the subset of the population who received treatment. Therefore, $Y|X = 1$ is the subset of $Y^{X=1}$ that we can observe. Some books, most notably Judea Pearl's text from 2009, prefer to denote $Y^{X=1}$ as $Y|do(X = 1)$, to emphasize that we have forced X to be a certain value for all individuals. The causal comparison thus becomes $Y|do(X = 1)$ vs $Y|do(X = 0)$ (Pearl, 2009).

Yet cases in which we can observe both counterfactuals for an individual are rare.¹ Many treatments are irreversible in some aspect; it is impossible to reasonably give a single subject every possible version. Suppose we wanted to examine the effect of long term smoking on cancer. A patient presents with a history of cigarette use and cancer. Based on the potential outcome definition of causality, we now need to contrast this history with what would have happened had the patient not smoked. This is impossible because we would need to go back in time and *ceteris paribus* prevent them from smoking. There is no other way to create a true test case and directly compare counterfactuals. Thus we must instead somehow simulate the unobserved counterfactual. Indeed, all methods of causal inference aim to minimize the distinction between the simulated and true counterfactual (Hernán & Robins, in progress).

¹One example is a crossover study in which the intervention group switches to being a control after a certain period of time and vice-versa for the control group. Information from such a design depends on assumptions regarding washout periods and other time trends.

1.2.2 Visualizing Causal Relationships with Directed Acyclic Graphs

Before we delve into methods used to make causal inferences, it is important to evaluate the context of the data generation. Blindly throwing every possible predictor into a gigantic model is not usually practical, efficient, or effective. Rather, we need to understand the connection between variables to know what to control for. Graphical representations are used to move towards this understanding.

Graphical representations of causal models were first developed by Sewall Wright in his 1921 book “Correlation and Causation”. However, his methods, known as *path analysis*, lacked a way to systematically incorporate statistical tools and tests.

Since Wright’s time, a variety of methods to illustrate causal relationships have been developed, including the *causal pie model* and *single world intervention graphs*, but many of these methods require use of specialized field knowledge. Instead, we will focus on *Directed Acyclic Graphs (DAG)*, accessible ways to visualize causal relationships that were developed primarily by two groups of researchers: Judea Pearl and colleagues at UCLA and Peter Spirtes and colleagues at CMU. (Pearl, 1988; Spirtes, 1996).

A typical DAG is given in Figure 1.2. It has two elements, event blocks and directed connecting arrows between them. An arrow connecting two events indicates a causal relationship between them, with the event at the origin being a direct cause of the event at the terminating end (Wright, 1921).

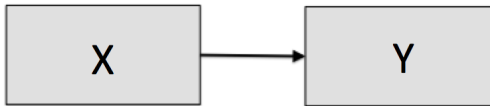


Figure 1.2: Simple Directed Acyclical Graph (DAG). The arrow indicates that the variable at its origin, X , is a direct cause of the variable at its terminating end, Y .

From this simple building block, complex relationships can be formed.

Three Variable Systems We begin by considering a three variable system (treatment X , outcome Y and other variable Z). This section details the three types of variables, defined by their relationships to the other two variables: *mediator*, *common cause*, and *collider* (Pearl, 2009).

Mediator A mediator is perhaps the simplest type of variable. X is a cause of Y , but acts through one or more mediator(s) Z . For example, consider the relationship between coffee (X) and cancer (Y). Coffee was once considered to be a cause of cancer. One of the primary reasons for this was that people who drank coffee in the 1970s generally smoked cigarettes as well. Yet it has been shown that smoking is a direct cause of cancer, while coffee is not a proven carcinogen. Suppose however, that

drinking coffee led to an increase in cigarette usage. Perhaps it brought people to cafes, where they would then join their friends for a smoke. Stopping these people from drinking coffee would then cause them to frequent cafes less often. This, in turn, would reduce their consumption of cigarettes and thus lower their risk of cancer. This relationship, shown in Figure 1.3, is clearly causal. However, it is not the act of drinking coffee, or anything having to do with the drink itself, that causes cancer. It is merely part of a series of events that will ultimately lead to this conclusion (Pearl, 1988; Spirtes, 1996).

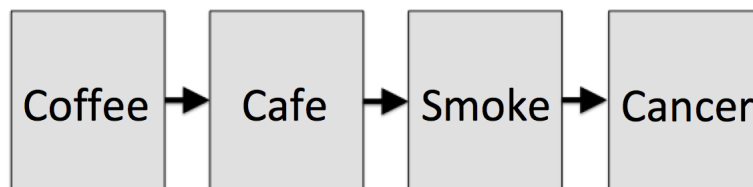


Figure 1.3: Mediator example. A mediator is an intermediary even in a causal chain linking the outcome and treatment. Mediators are active in their natural state. Therefore there is an observable association between the outcome and treatment, unless the mediator is conditioned on. In this example, Coffee (X) is a cause of Cancer, but only through the mediators Cafe and Smoke.

Common Cause The second type of variable is a common cause. Common causes arise when there exists a variable Z that has a causal effect on both X and Y . When students are told that not every association is causal, it is often because of the existence of common causes. These are problematic because even if X and Y have no causal relation, the presence of Z gives them an association. This association makes it appear as if X changes with Y , but that is not true.

Consider again the example of the relationship between coffee, smoking and cancer. Nowadays, smoking is much less common in cafes. In fact the practice has decreased so much that coffee is no longer a cause of smoking. If anything, the relationship may have reversed as smoking is thought to cause insomnia, resulting in smokers needing more caffeine to keep them awake during the day. Thus, smoking is now a common cause of coffee drinking and cancer risk. Figure 1.4 illustrates this relationship.

Note that there is still an observed association between $\{Coffee\}$ and $\{Cancer\}$, but there is no continuous directed path and thus no causal relationship. Even if we were to intervene on $\{Coffee\}$, if we didn't also address their smoking habit we wouldn't see a change in $\{Cancer\}$. Thus common causes produce association but not causation (Pearl, 1988; Spirtes, 1996).

Collider (Common Effect) The final type of variable is the *collider*, sometimes called a *common effect*. Here X and Y both have causal effects on some event C , but are not related in any other way. In our example, suppose we incidentally only select patients who have yellow teeth, which is an effect of both coffee and smoking (with

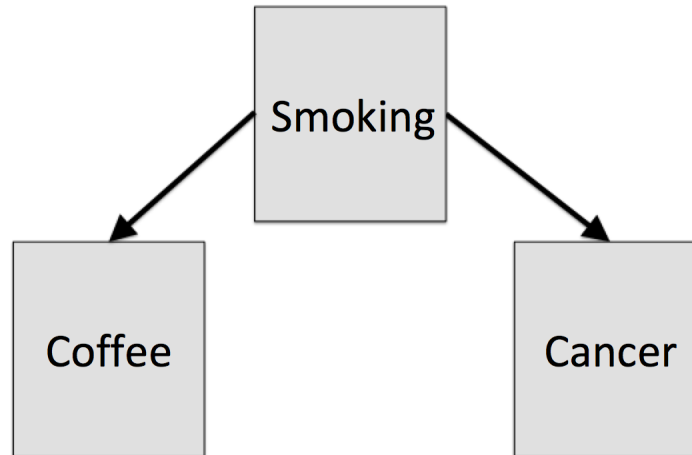


Figure 1.4: Common cause example. A common cause is a cause of both treatment and outcome. It is active in its natural state. Therefore there is a spurious association between the treatment (Coffee) and outcome (Cancer) in this graph, unless the common cause (Smoking) is conditioned on.

smoking, in turn, causing cancer). Other than this link, we shall assume there is no other relationship between coffee and cancer. Figure 1.5 shows a possible DAG for this relationship. Note that this diagram is different from those depicted before, since without conditioning there is no observed association between $\{Coffee\}$ and $\{Cancer\}$.

However, if we were to condition on $\{Yellow\ Teeth\}$, there would be information to be found. Say we note that our smoker has yellow teeth and is not a coffee drinker. If this diagram is true, and there are no other unknown and unmeasured causes of yellow teeth, we can guess that they smoke and will likely have cancer. Conversely, if a person has white teeth, and is not a coffee drinker, they probably do not smoke as well. Thus, controlling for a common effect creates an artificial association between the two causes. This is why “kitchen sink” methods of simply controlling for all available covariates is problematic. Treating common effects as confounders will result in a false positive.

Summary Combinations of these three formations can be combined to create complex systems that describe most causal relationships. Only the first (mediator) depicts causality, while the second (common cause) will still appear as an association. The third (collider) can induce an artificial association if not properly controlled for. Now that we have our basic building blocks, we can use them to explore causation (Spirtes, 1996).

Active Paths In order to determine causality from a DAG there must be at least one directed, active path between the cause and effect. A *directed path* means there exists an unbroken string of mediators between the cause and effect, such as $A \rightarrow B \rightarrow C \rightarrow \dots \rightarrow E$. A directed path is an inherent condition of a relationship. There is

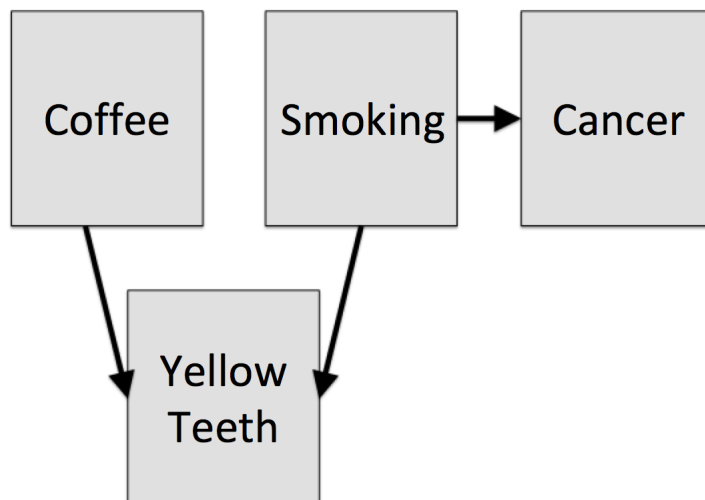


Figure 1.5: Collider example. A collider is a variable that is an effect of both the treatment and outcome. It is inactive in its natural state, meaning there is no association between the treatment (Coffee) and outcome (Cancer, through smoking) in this graph. Conditioning on a collider (Yellow Teeth), however, will activate it and induce a spurious association.

no way to change its state as it is either true or not. An *active path* on the other hand depends on exterior forces. It is a state rather than a trait. Any path made up of only mediators or common effects, and thus produces an association between X and Y , is active in its natural state. If it contains one or more colliders, it is inactive. However, as soon as you condition on an event, its state switches. Thus if you condition on any variable on a naturally active path it becomes inactive and if you condition on every collider on a naturally inactive path it becomes active. We can use these rules to identify potential causal relationships (Pearl, 2009).

1.2.3 D-separation

When analyzing data it is often useful to determine how two variables are related, if at all. This is one of the primary uses of DAGs. It can also be important to understand the consequences of controlling for a set of variables, Z , referred to as the *conditioning set*. We begin, as always, with an example (see Figure 1.6):

This is the most complex DAG we have seen, but can still be broken down into 3-variable systems that correspond to the ones we know how to deal with. For this example, we are interested in the relationship between X and Y . Notice that the “A” path (red) has a common cause, the “B” path (green) has a mediator, and the “C” path (blue) has a collider and a descendant of the collider ($C1$). A *descendant* is any causal effect that stems from a collider. In other words, the collider C is also a mediator between X and $C1$. We will see why these are important later. For now, just remember that the path through B establishes causation of Y by X , the path

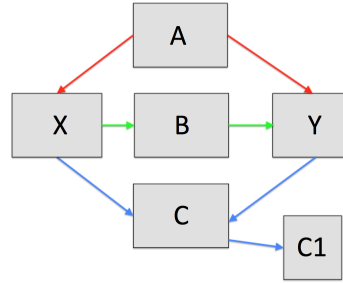


Figure 1.6: D-separation example containing a common cause (red path), mediator (green path) and common effect (blue path)

through A only association, and the path through C displays neither.

Now let us suppose we decide to control for some variables in the conditioning set, Z . In analysis, variables in Z are often those that are added as a predictor to a regression model or stratified so that only one level is observed at any given time. If a variable is in Z , its state is switched from active to inactive or vice versa. Thus a common cause or mediator will become inactive while a collider will become active. Note that conditioning on a descendant will have an equivalent effect as conditioning on the collider it stems from. Therefore, adding $C1$ to Z will have the effect of turning C active.

Once we have our conditioning set, we can determine the *dependence separation* of X and Y , also known as D-separation. X and Y are *D-separated* by a path if, for a given Z , there exists at least one variable on that path that is inactive. If all variables are active, the path is said to be *D-connected*. To illustrate this idea we return to our example. If $Z = \{\emptyset\}$, X and Y are D-connected by the path through A and the path through B . They are D-separated by the path through C . But, if A is an element of Z , the path through A D-separates X and Y . If C or $C1$ are elements of Z , the path through C D-connects X and Y (Spirtes, 1996).

But why do we care if X and Y are D-separated by a particular path, given Z ? The answer is that D-separation is a useful method to test the accuracy of a particular DAG. Recall that most DAGs are created based on observations or expert knowledge. They are usually proposed relationships that need to be confirmed in some way. D-separation offers us a way to do so. Say X and Y are correlated. This means that there exists at least one active path between them. Hopefully our DAG reflects this. If it does, we can then begin to add variables to the conditioning sets such that no active paths are left. We should then observe no association between X and Y and experiment with different combinations of variables in Z , confirming our predictions about whether or not we observe association.

Caution! While D-separation is a useful tool in determining associational pathways between X and Y , it *does not imply causation*. An active path is a necessary, but not sufficient requirement for causality. We still need directionality between X and Y . This can only be discovered through expert knowledge or manipulation (Spirtes, 1996).

DAGs in Use

In practice, we are not provided a pre-made causal graph. Rather, we must make one ourselves. This example expounds on the potential causes of cancer by starting with an actual study that suggested coffee causes cancer, based purely on an association (Figure 1.7). From now on we will indicate the relationship of study with a dotted gray arrow behind a question mark. All other relationships in our proposed DAG are solid black arrows. This is to emphasize the study relationship and consistently remind ourselves of the study goal.



Figure 1.7: DAGs in use. Dotted gray arrows with question marks indicate the (unknown) relationship of study. In this DAG, coffee is the treatment while cancer is the outcome.

The coffee/cancer relationship is a stereotypical example of the aphorism that “association does not equal causation”. But why do they appear related? Discounting random chance, there are several possible reasons, each associated with a different graph but all including the presence of a third variable. Note that we are assuming a temporal relationship between coffee and cancer in that heavy coffee consumption occurs throughout a patient’s life, even before they have cancer. Given that coffee occurs before cancer, cancer cannot possibly be a cause of heavy coffee consumption.

We start by exploring the literature to find potentially relevant causes of cancer. An example of potential diagrams can be found in Figure 1.8. The first and most obvious example is smoking. Therefore we will use smoking as our third variable. Note again the temporal aspect of the smoking/cancer relationship. We assume that lifelong smoking does its damage before the development of cancer. Thus we can avoid issues of reverse causality and assume cancer cannot possibly be a cause of smoking.

Our first step is to think of possible DAGs that could exist (Figure 1.8). We can then use D-separation to test their accuracy (see Table 1.1).

Conclusion

In order for a binary treatment variable A to have a causal effect on Y , $Y^{A=1} \neq Y^{A=0}$ and there must be an active, directed path from A to Y . Yet, as alluded to before, it is extremely rare to be able to observe both counterfactual outcomes $Y^{A=1}$ and $Y^{A=0}$. In fact, the term “counterfactual” stems from the idea that typically only one of these outcomes is observable. Usually, treatments are irreversible and patients can be in only one of the treatment or control groups. This is the “fundamental problem of causal inference” (Holland, 1986).

All of our techniques and experimental designs are created in order to approximate a setting in which we observe both counterfactual outcomes. At this point it is helpful

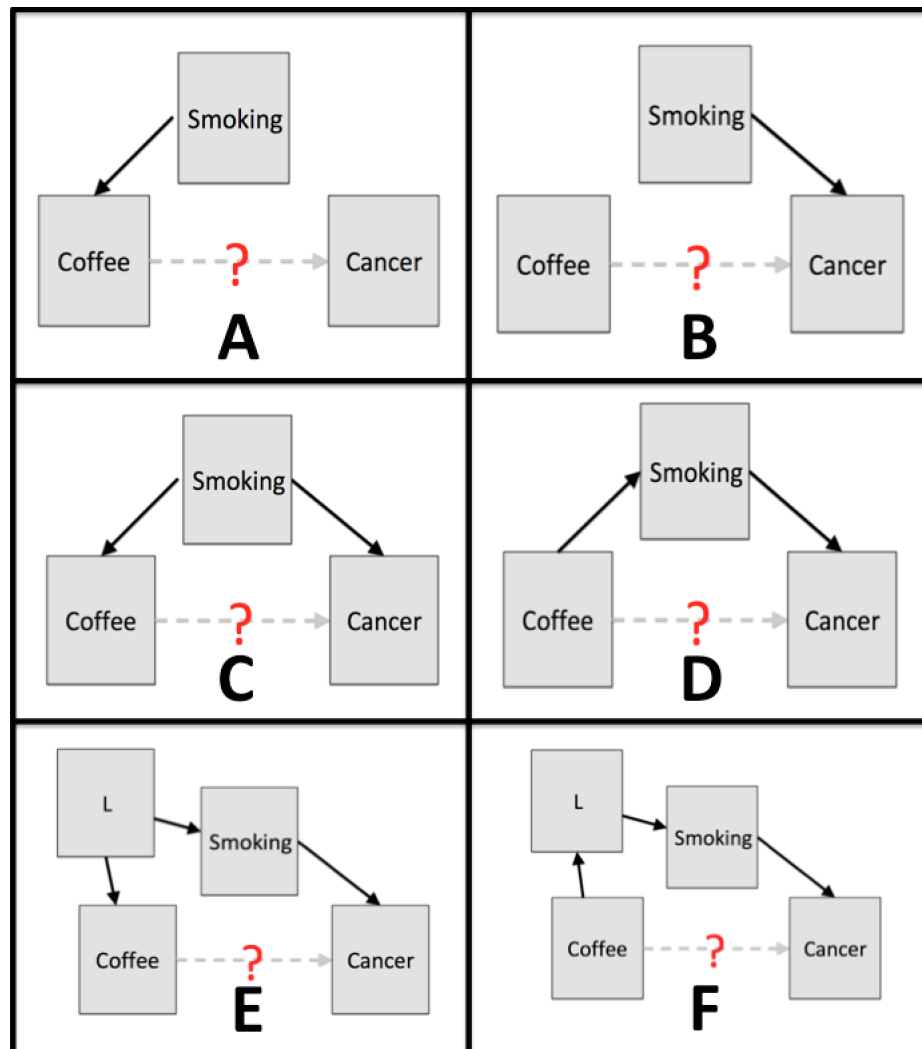


Figure 1.8: DAGs illustrating the potential relationships between coffee, cancer and smoking. L represents any other lurking variables that may lie on the causal chain between the treatment (coffee) and effect (cancer).

DAG	Association between Coffee and Cancer	
	With $Z = \emptyset$	After Condition on Smoking
A	No	No
B	No	No
C	Yes	No
D	Yes	No
E	Yes	No
F	Yes	No

Table 1.1: Testing Potential DAGs. Suppose there is no direct causal effect between coffee and cancer (if there was, there would always be an observed association). There will still be an observed association in DAG C through F. However, only D and F depict a true causal effect. By only controlling for smoking, they cannot be distinguished from their non-causal counterparts.

to examine the simplest and most ubiquitous causal inference method: the randomized controlled trial (RCT).

Chapter 2

Estimating Causal Effects

2.1 Using RCT Data

Most people studying science have some idea of the difference between association and causation. As Barnard (1982) once noted, “that association is not causation is perhaps the first thing that must be said”. Yet the idea of “no causation without manipulation” has taken an undue influence on causal inference (Holland, 1986).¹ Some professionals take randomization to be the primary indicator of reliability, limiting the scope of their knowledge and potentially leading to false conclusions if they improperly trust results from poorly designed studies simply because there is an element of randomization (Rubin, 2008). This section will illustrate two points. First, it will detail the fundamentals of experimental design and how randomization of treatment reduces confounding. Then, we shall illustrate necessary assumptions for causal inference in observational data. Finally, we will conclude with a brief overview of common families of methods.

Recall the notation defined in Section 2. Let X be a dichotomous treatment (1 = treated, 0 = untreated) and Y be an outcome variable of interest. To keep things simple, suppose Y is dichotomous with $Y=1$ resulting in death and $Y=0$ representing survival. $Y^{X=x}$ represents the outcome given the level of treatment. For X to have a causal effect on Y , $E(Y^{X=1}) \neq E(Y^{X=0})$. Furthermore, there must be a directed, active path Z that connects X to Y .

Randomized Controlled Trial: Why it works

A typical randomized controlled trial takes a sample from a population and randomizes participants to a treatment or control group. The treated group has their counterfactual outcome $Y^{X=1}$ known but $Y^{X=0}$ unobserved, while the opposite is true for the control group.

The premise of the design is that each respective group’s observable counter-

¹Interestingly Holland and Rubin, the authors who initially coined the term “no causation without manipulation”, later regretted the success of their slogan (Holland, 1986). As some of the biggest and most prolific proponents of causal inference, they have since spent a lifetime moving beyond this statement.

factual outcomes are interchangeable. This is our first identification assumption: *Exchangeability*. Formally, it is noted as $E[Y^{X=1}|X = 1] = E[Y^{X=1}|X = 0]$ & $E[Y^{X=0}|X = 1] = E[Y^{X=0}|X = 0]$. More concisely: $E[Y^{X=a}] \perp\!\!\!\perp X$. In less quantitative terms, this suggests that had we switched the groups (i.e., the treatment group had been the control and the control group had received treatment), we would have observed the same result. To understand why this is so important, recall our fundamental method of determining causality: comparing counterfactuals. In this experimental design, the control group simulates the unobserved counterfactual of the treatment group. This is only possible if all relevant non-experimental aspects of the control group are identical to those of the treatment group. Because of this assumption, we can compare counterfactual outcomes and determine causality (Rubin, 1974).

2.1.1 Confounders: A Lurking Danger

There is a natural question that arises from such a design: how can we determine if exchangeability holds?

We shall illustrate an explanation using DAGs. For our counterfactuals to be adequately compared, we need the population receiving treatment to be comparable (exchangeable) to those in the control group. Thus there can be no variable or set of variables, L , that have a causal effect on treatment level and outcome. In DAG terms, we need to rule out the possibility that X and Y have a common cause (Figure 2.1).

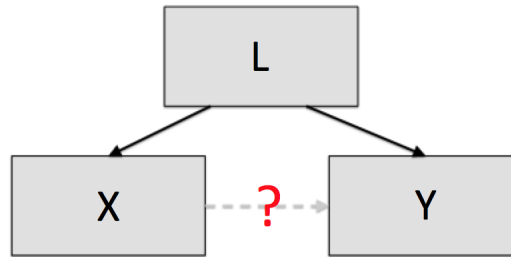


Figure 2.1: Confounder example. Here we see that L is a confounder of X and Y . While L remains uncontrolled, it is impossible to make a determination on the relationship between X and Y , if one even exists.

Any set of variables, L , that serve as a common cause linking X and Y are known as *confounders*². L opens an active causal pathway between X and Y , offering a plausible non-causal pathway that may explain any observed association (Spirtes, 1996). With this active path in place, it is impossible to make any determinations on the relationship directly connecting X and Y (the path with the question mark in Figure 2.1). No matter what direction the arrow points, if it exists at all, there will always be an observed association. If we neglect to fully account for L , we may mistakenly believe this association implies causality. Methods to mitigate confounding will be described in Section 3.

²Some texts refer to these as “other factors” or “lurking variables”. See Appendix for more detailed information.

Randomized controlled trials are generally considered to be robust against confounding, at least relative to observational studies. So what makes an experiment so special? The answer lies in randomization.

The Beauty of Randomization

The easiest way to avoid confounders is to randomly assign treatments. Then, there is a known probabilistic chance that an individual receives a certain treatment and on average there will be no common cause, L , of the treatment and outcome. In DAG terms, randomization will, on average, break the causal chain between L and X (Figure 2.2). This effectively eliminates L as a confounder and cause of X (Spirtes, 1996).

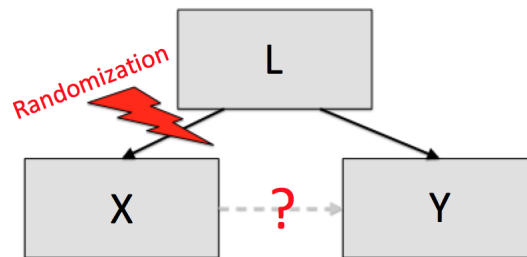


Figure 2.2: Randomization works by breaking the connection between assignment of treatment and any external covariate. Thus, it eliminates potential confounders and minimizes the risk of bias.

Correct On Average Simple random assignment of treatment, while an important breakthrough in design, is not infallible. In truth, it is only correct *on average*. Because it relies on randomized allocation it may be possible for especially skewed or unbalanced treatment groups to occur, particularly for small studies. This is why it is important to gather large sample sizes, as the probability of significant deviation from expected balance is inversely proportional to the number of participants (Rubin, 1974). Research in this area led Rubin to later muse whether “for gold standard answers, complete randomization is good enough, except for point estimation in very large experiments” (Rubin, 2008).

All of this means that randomization is an important tool to mitigate confounding, but is not itself a sufficient solution. Further experimental considerations will be discussed in Section 2.1.3.

2.1.2 Effect Modification

There generally is no such thing as *the* causal effect of X on Y . Rather, there is an effect of X on Y given a certain study population. If the causal effect varies across some characteristics of the population, and there is no underlying structure causing confounding, these characteristics are called effect modifiers. Race, sex and other inherent traits are common effect modifiers.

Mathematically, effect modifiers, M , satisfy the following criteria:

1. $M \perp\!\!\!\perp X$
2. $E[Y^a \mid M = 1] \neq E[Y^a \mid M = 0]$,

where a is a constant representing the assigned treatment.

Although they may seem similar, effect modification is distinct from confounding. Confounding is a problem to be addressed, failing to account for it results in a biased answer. Effect modification is a structural property of the data where different sub-populations have slightly different effects. Failing to account for them simply results in an average over these distinct groups and perhaps losing some specificity. For instance, say we decided to study the effect of birth control on fertility. It is possible for the study to take a sample from the general population, men included. We could then give everyone in the treatment group birth control pills and count the number of children they have (assuming their partner is also not on birth control). This study would produce a true result, as the observed difference is the rate in the general population. However, due to effect modification, it may not give the most clinically meaningful one. The presence of an effect modifier does not bias the conclusion, just lessens the precision and, potentially, its usefulness (Hernán & Robins, in progress).

2.1.3 Selection Bias

Selection bias is a broad term that has different definitions depending on a researcher's background or discipline. Roughly speaking, it refers to the issue of having a study population that is not representative of the source population. Note the distinction this creates between a confounder and selection bias. A confounder is a lurking variable that induces bias through its association with both the treatment and outcome. Put simply, it is the imbalance in the covariate distribution across treatment groups. Selection bias, as opposed to confounding, is not an issue of treatment assignment balance. This means the randomization of treatment groups cannot generally solve selection bias in most real world settings. Thus, investigators must be especially alert to issues that may induce selection bias.

Some people consider everything that has to do with participants selection to fall under the umbrella of selection bias. Other authors, notably Hernán and Robins, take a slightly more nuanced view and consider selection bias to consist of only common effects that, after being improperly conditioned on, lead to an artificially induced correlation. They consider all other forms of “selection bias” really just issues of generalizability (Hernán & Robins, in progress). Unfortunately, there is no clear consensus on the definition of selection bias. We shall draw on Hernán and Robins' definition, as it is the most comprehensive definition in a book dedicated to causal inference and separates biases by how to address them.

In general, selection bias can be illustrated by adding a new event to our DAG, such as in Figure 2.3. This event is a common cause of both treatment and outcome and thus creates spurious association if conditioned on.

There are several different forms of selection bias. Some include:

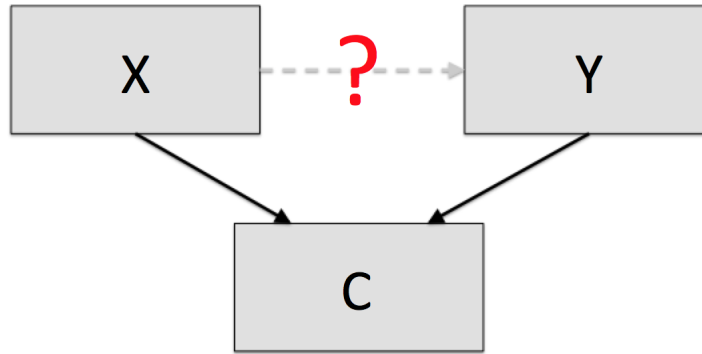


Figure 2.3: Selection bias example. Selection bias arises from improperly conditioning on a common effect such that a spurious association is induced. Examples include loss to follow up and noncompliance

- *Loss to Follow Up*: Experiments can be long, time consuming and offer little immediate benefit. Patients are often lost over the course of a study. Moreover, the sickest, highest risk, and least compliant patients often drop out at a higher rate. This can bias the final results and can potentially lead to incorrect conclusions.

It is possible to mitigate the effect of loss to follow up during analysis. We can assign each individual a probability of dropping out (based on an educated guess), then use the inverse of this probability as an additional weighing covariate in the style of inverse propensity weighting. This will give a higher weight to the results from people similar to those who dropped out and account for their loss. This weighting method will be covered in more depth in weighting section (Hernán & Robins, in progress).

- *Noncompliance*: Study participants may not follow the treatment they have been assigned. A person who follows instructions is said to *adhere* to them, whereas a person who does not is said to be *noncompliant*. In the case of a RCT, this can disrupt even the most meticulously planned experiments. Researchers often have no idea if a patient actually complies with their instructions and have no way to control for it. Thus, clinical trials are often classified as *intention to treat* (ITT). These studies only consider the level of treatment *assigned*, disregarding what the patient actually did.

While it seems odd to neglect the actual treatment received, an experiment should always be viewed in terms of its purpose. If patients in the treatment group need to inject their medicine twice a day, but those in the control do not, the treatment group may be less compliant. Therefore, the effect of the medication may be underreported.

ITT studies are designed to address the question of whether or not a physician should prescribe a drug. In practice, some patients will be noncompliant. If the intention was to determine the biological mechanism through which the drug works, then knowing the received treatment would be necessary. But in this situation, when we are trying to predict the future results of a patient being

prescribed the drug, intention to treat may be more informative (Pearl, 2009).

2.1.4 Sampling Bias

Anything that occurs before the assignment of treatments that causes the study population to differ from the target population is *sampling bias*. If left uncorrected, study findings may not generalize to the larger population and have the intended outcome.

Some authors do not consider sampling problems as true issues of causality. They point out that a well defined causal effect estimated from an unrepresentative sample is still a causal effect, even if it holds true for only a subset of the population. The issue is having the causal effect you think you have, not correctly estimating one (Hernán & Robins, in progress). As such, this topic is not covered in many books on causal inference. However, estimating the correct effect is important and every study should be analyzed for sampling bias.

A classic example of sampling bias is from drug testing in animals. It used to be that only male mice were used to test drugs, as researchers prioritized consistency in treatment by avoiding female hormones. Eventually though, people began to recognize that this is bad practice, as female humans also use the drugs (Beery & Zucker, 2011). The findings were not generalizable to the entire population, only to males sharing similar characteristics to those studied.

Other common sources of sampling bias come from self selection, as study participants are often not representative of the population as a whole. This is sometimes referred to as healthy or volunteer bias, as studies often neglect the sickest patients who are not able to participate. One famous case of sampling bias came from the hormone replacement therapy (HRT) trials to treat heart disease. Here, estimates were taken from the Nurses' Health Study, whose population is primarily younger, healthier women. The eventual results were misleading, as the treatment is effective only immediately following menopause. The study happened to take place with the ideal treatment population and the results did not generalize as well to the larger population. This example is discussed further in the applications section (Lobo, 2017).

Sampling a subset of the population is not inherently harmful. In fact, it is usually necessary. Bias occurs when the sampled subset differs from the target population by some factor that is associated with the outcome. In the HRT example, age was a confounder for the treatment effect, thus limiting the study to younger women led to bias. If we had instead only sampled left handed women, and left handedness was neither an effect modifier or confounder, then there would be no bias.

Equipoise Although the claim of a causal effect of smoking on health is widely accepted to be true, but there has never been a randomized controlled trial to confirm the results. This is largely due to the fact that it would be ethically dubious to intentionally expose the treatment group to harm in the form of cigarette smoke. Even if a cigarette study were to be approved, what patient would sign up for a treatment that is likely to harm them?

In response to this problem, investigators have come up with an ideal standard to

strive for: *equipoise*. Equipoise is defined as “the point where a rational, informed person has no preference between two (or more) available treatments” (Lilford & Jackson, 1995).

Conclusion

Randomized controlled trials are essential tools of causal inference, but are not infallible. Every consideration should be made to ensure a RCT is as close to ideal as possible. RCTs are not the only method generating data that can be used to draw causal conclusions, however most observational methods are judged by how closely they resemble an ideal randomized experiment. Thus, it is important to develop a solid understanding of RCT design before exploring other approaches.

2.2 Using Observational Data

There has historically been a great deal of controversy about the role of observational data in causal inference. Some believe studies based on such data are inherently anecdotal and uninformative and they bristle at the thought of using their conclusions (Holland, 1986). Others believe that, while not as reliable as data generated from experiments, observational data do have something to offer and, with so much found data available, it is irresponsible of us to neglect them (Rubin, 1974). We shall take the second view, detailing several identification conditions that must be satisfied before applying causal interpretation and briefly summarizing several common estimation methods.

2.2.1 Identification Assumptions

There are three fundamental considerations that must be made before analysis to ensure that a causal effect can be estimated.

Positivity For treatment X to have a causal effect on Y , an individual must have a positive, non-zero probability of receiving the treatment and control. If they could not receive a different level of X , we could not possibly intervene to affect the state of Y .

The most common violation of positivity is lack of an adequate comparison. If all subjects receive the same level of treatment, it is impossible to say if it is the origin of an observed change. There’s simply no counterfactual estimate to use for the requisite comparison. For example, if we wanted to observe the impact of smoking on the development of cancer, we need people who have and have not smoked. Otherwise, it is possible that everyone develops cancer at the observed rate and we cannot isolate a distinct rate in smokers.

Another consequence of positivity is that inherent, unchangeable traits cannot be causal. This suggests qualities such as race, sex, place of birth and parental characteristics cannot cause something to happen. To understand why, we need to recall the general purpose of causal inference: intervention. We seek causal chains

so that we can affect the world around us. Determining that males are generally more obese than females does not help us address the disease. But recognizing that being male is associated with eating a greater amount of calories, which then causes obesity is an important finding. Positivity helps to justify our assertion that any observed change is real, and ensures that any causal relationship we find gives us the information we seek (Hernán & Robins, in progress).

Consistency Before beginning an analysis, it is important that the treatment be *well defined* so that there is no ambiguity as to which group a patient may belong. For instance, if a researcher were examining the effect of smoking on cancer, they need to define what exactly constitutes a smoker. Are past or occasional smokers included? How about people who smoke e-cigarettes, do they fit either category? A better definition would be those who currently smoke at least a pack of cigarettes (20) a week on average, with e-cigarette users not being eligible for the study. This ensures that there is no confusion or misassignment of groups.

Yet having well defined groups is not enough. Ideally, all subjects in a particular group will receive a *consistent treatment*. Otherwise, some people in the treatment group may receive an inferior form of care and have a negative outcome, whereas if they had received the same level of care as everyone else they would have a similar outcome. For example, even if the above definition for smokers is precise in its classifications, the control group may not have consistent treatment. People who smoke half a pack of cigarettes a week (10) are very clearly in the control group, but they probably have distinctly elevated cancer risks from those who have never smoked before. This will raise the average cancer risk in the control group and may hide a true causal effect (Hernán & Robins, in progress).

Exchangeability Although we have already discussed exchangeability, we shall briefly explore some of its subtleties. Exchangeability states that, all else being equal, the counterfactuals of the treatment and control group are comparable. The statement, “all else being equal”, is very important. We would not reasonably expect someone who smokes sporadically to have comparable outcomes with another who smokes a pack of cigarettes a day. Thus our formal definition is modified to be $E(Y^{A=1}|A = 1, L) = E(Y^{A=1}|A = 0, L)$. Under the assumption that all measured covariates remain constant, we want the expected outcome under treatment to be the same regardless of actual treatment assigned.

It may seem that exchangeability and consistency are very closely related. This is correct, as they both essentially ensure that observable counterfactuals are comparable. However consistency deals with comparability within groups, whereas exchangeability ensures inter-group comparability, assuming all other necessary covariates are the same.

Of course it is generally impossible to determine if all relevant covariates are controlled for. It is up to researchers to utilize their expert knowledge of the subject before proceeding. Worse yet, we are not able to empirically determine if our assumption is valid, until another study is published that disproves our claim. While it is possible

to test a study's sensitivity to lurking variables, it is impossible to be truly free from them (Hernán & Robins, in progress).

2.2.2 Validity of Resulting Causal Conclusions

Many journals and institutions have begun releasing guidelines to help their members evaluate studies utilizing observational data. While some go into greater depth than others, there are several common themes. For instance, most guidelines suggest comparing the observational study to an idealized randomized trial. Below are some examples of criteria to evaluate observational studies.

Early definitions: Bradford Hill Sir Bradford Hill was an English statistician who proposed one of the first sets of criteria to prove causation. This list of ten qualifications was a groundbreaking attempt to understand a concept poorly understood by scientists of the time and can be seen in Table 2.1 (Hill, 1965).

GRADE: A Standardized Evaluation Criteria The “Grades of Recommendation, Assessment, Development, and Evaluation” (GRADE) criteria were created to understand the risk of bias from a study. Developed by a large group of health researchers, the results were published in the *Journal of Clinical Epidemiology* (Guyatt, 2010). They listed four limitations of observational design. The first is a failure to develop and apply eligibility criteria. This includes experimental elements such as use of control population, under/over matching and selection of exposed in cohort studies from different populations. The second is a failure to gather accurate measurements of exposure or outcome. Next, they highlight issues of confounding and, finally, incomplete follow up.

The actual score and criteria of GRADE are the subject of much debate. Criticism focuses on the opaqueness and inconsistency of scoring and heavy bias in favor of randomized trials (Michelle Irving & Anstey, 2017).

Modern Considerations There have been several recent publications designed to help people evaluate the effectiveness of observational studies. They are notably looser than Sir Bradford Hill's, with far fewer criteria. For instance, Roger Peng's list in the *American Journal of Epidemiology* consists of three questions (Peng, Dominici, & Zeger, 2006):

1. What are the actions or exposure levels?
2. Is there an adequate comparison group?
3. How closely does it approximate an idealized randomized study?

Thomas Frieden (2017) takes an even more holistic approach, suggesting that every situation is unique and flawed. He believes that conclusions should be made based on evidence from multiple sources and a solid understanding of causality is more important than a set of evaluation criteria.

1. Temporality	Did exposure happen before the onset of the disease?
2. Strength of Association	Is the measure of association between the exposure and outcome strong?
3. Dose-Response Relationship/Biological Gradient	Do people with a higher level of exposure have a higher risk of the outcome than people with a lower level of exposure?
4. Cessation	Does stopping the exposure reduce the risk of the outcome?
5. Specificity	Are the exposure and outcome both narrowly defined rather than general concepts?
6. Theoretical Plausibility	Is there a reasonable biological explanation for why the exposure might cause the outcome?
7. Consistency	Has a potentially causal relationship between the exposure and outcome been observed elsewhere?
8. Coherence	Is there a causal relationship between the exposure and outcome congruent with other knowledge?
9. Consideration of Alternate Explanations	Are there reasons the apparent causal relationship might not actually be causal?
10. Experimentation	Is it ethical to conduct an experimental study? Has experimental testing confirmed a causal relationship?

Table 2.1: Sir Bradford Hill's criteria to determine causality (1965).

2.3 Methods to Determine Causal Effects

There are a variety of methods to deal with causal inference. The vast majority of these approaches boil down to one simple principle: allow for the best comparison of counterfactuals. All of these methods, whether they manipulate the outcomes or

minimize the differences in covariates, can and should be thought of as applications of this idea. But methods of causal inference share another characteristic. In general, all approaches — even randomized trials — rely on certain untestable assumptions. It is up to the investigator to use their expert knowledge to determine the validity of their counterfactual comparison. Methods can reduce the burden of these assumptions up to a point but can never truly satisfy them.

This section is primarily meant to introduce several families of methods in an effort to showcase their underlying principles and provide a reference for more complex approaches.

2.3.1 Stratification

Stratification is the practice of examining sub-populations separately, and then comparing their outcomes. It is usually used to isolate populations of interest or to deal with effect modification. It can be especially useful when used on dichotomous or discrete variables with few levels.

In our example of the relationship between coffee and cancer, many of the initial studies noted that males had a significantly greater association than females. To find this information, they must have grouped and compared the study participants by sex. After observing that the results in the two groups differed³, the researchers were able to confirm the presence of sex as an effect modifier (Hernán & Robins, in progress; Pearl, 2009).

2.3.2 Confounder Matching

Another way to ensure balanced counterfactuals is to directly *match* each observation in the treatment group to one or more analogous observations in the control group. By ensuring that every individual has a match, we can be assured the positivity condition (that every level of treatment is possible) holds and that there is a good effect measure for every set of covariates. Once we have a matched population, we can choose an appropriate analytical technique, whether it be treating each matched pair as a stratified subpopulation or creating a model based on the subset of the total population that has an effective match.

Exact Match

The gold standard in matching is an exact match on a broad set of covariates. Suppose Z_T is a set of all covariates that D-separate a treatment, X , and outcome, Y . For any treated individual, u_t , we can define an exact match, u_c , as any individual whose Z_T is exactly equivalent to the treated individual. In this setting we have a perfect counterfactual, as the subjects will only differ by X and, potentially, Y .

Of course, exact matches are uncommon. Even if we match on all observed D-separating covariates, denoted Z_o , it is generally impossible to say if all variables

³Note that groups may differ by additive, multiplicative or other comparisons. The choice of metric depends on the situation and intended finding.

are accounted for. If we are wrong, and $Z_o \subset Z_T$, we may still have a confounder misrepresenting our causal relationship. This problem is not unique, recall the discussion of lurking confounders, but should always be considered when evaluating a supposed exact match.

Another major limitation to matching exactly is lack of data. It is often very hard to find individuals with identical Z_o , let alone Z_T . Trying to restrict a population to only individuals with a match may cause the sample size to be reduced so much that all power is lost. Moreover, we may end up limiting our analysis to include only individuals with common covariate sets, as people with distinctive features may not have a match. This could create an artificial selection bias that under-represents patients who are already in the minority.

Despite these limitations, exact matching on a broad set of possible confounders is the ideal standard. All other methods strive to replicate exact matching, while slightly loosening the restrictions to mitigate the obvious flaws (Roy, Rudin, Volfovsky, & Wang, 2018).

Coarsened Exact Matching

One of the most common issues with exact matching is trying to match on continuous data. This is because it can be rare, or even impossible, to correctly identify individuals with the same exact levels. While it is possible to perform a “fuzzy” or optimal match, where matches are found within a certain manually-defined distance, a more common method is to “coarsen” the variable, or break it into discrete categories. Once the variable is coarsened, we can match exactly. There could still be residual confounding, but likely only second order bias (Roy et al., 2018).

An example of this would be tax brackets. Income is a continuous variable that would be nearly impossible to match on exactly. But, as of the most recent tax bill, people are divided into 7 discrete categories according to their income. Within each bracket individuals’ tax rate are the same. Now, rather than trying to match on a person’s true income, we can match on their tax bracket.

Unfortunately, while coarsened exact matching allows us to preserve dimensionality, this can be a curse as well as a blessing. More brackets mean smaller categories, and in sparse datasets with large number of covariates, it may be impossible to find exact matches.

2.3.3 Weighting

Suppose researchers want to obtain a reliable estimate of the smoking population. However, most of the study participants who volunteered to be interviewed were elderly and less likely to smoke. Thus, the initial sample was biased and had a lower proportion of smokers than we expect.

In order to correct for this issue, the researchers decided to *weight* the sample. To do this, they first stratified the results by age group, noticing that each group had markedly different smoking habits. They then calculated the proportion of people in that age group, took the inverse and multiplied it by their rates. The inverse of the

proportion was used so that underrepresented individuals were counted more heavily than people who were already accounted. The equation will be:

$$\widehat{Smoking\ Population} = \sum_{i \in \text{age groups}} \left(\frac{Total\ Population_i}{Sample\ Population} \right) (ave\ \# \text{ of cigarettes})_i,$$

Where $population_i$ is the number of people in the i^{th} age bracket and $ave\ \# \text{ of cigarettes}_i$ is the average number of cigarettes consumed by people in i^{th} age bracket. By first stratifying, then adjusting based on population characteristics, the researcher was able to achieve an accurate estimation of the population. This is just one example of weights. Other weights, using propensity scores or other methods can also be used depending on the situation.

How is this related to causal inference? Oftentimes, counterfactuals do not have the same characteristic of the population distribution. Thus, they are not truly exchangeable and any comparison may be inaccurate. Weighting works to make the populations more similar by emphasizing individuals underrepresented in one counterfactual and limiting the influence of those who may not appear in both. This can make an especially important and meaningful difference in small sample sizes, though it can be hard to estimate weights in such situations (Hernán & Robins, in progress).

2.3.4 Regression Control

Regression models are useful methods in causal inference. Consider the interpretation of a multiple regression model of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon_i,$$

where $\{x_1, \dots, x_p\}$ are a set of predictors, $\{\beta_1 x_1, \dots, \beta_p x_p\}$ are their coefficients, and ϵ_i is an error term. Suppose we let x_1 represent our treatment variable, taking a value of 1 if the subject received treatment and 0 otherwise. β_1 can be interpreted as the change in the outcome, Δy , under treatment for a given set of covariates. This is precisely what we are looking for. By definition, we are comparing counterfactual outcomes in analogous individuals.

Modeling approaches have an advantage in that, unlike other techniques, they allow for flexible incorporation of continuous levels of treatment, as opposed to only dichotomous or categorical. If we are interested in a dosage response or other continuous treatment, modeling approaches become very important. There is also a large and rigorous body of work behind them, with techniques developed to deal with many different types of situations. This is especially true when we consider that causal inference methods are not restricted to linear regression models. Logistic, Poisson, or any other family of models work just as well under the correct circumstances (Hernán & Robins, in progress; Rubin, 1974).

2.3.5 Propensity Scores

Propensity scores were first described by Rosenbaum and Rubin (Rosenbaum & Rubin, 1983, see An & Ding, 2017 for a recent review). The idea is to balance treatment groups by using measured covariates to predict the likelihood that an individual will receive treatment. Mathematically,

$$\text{Propensity} = \Pr(X = 1 | L = l),$$

where $\Pr(X=1)$ is the probability that an individual receives treatment and L are all measured covariates. Logistic regression of the form,

$$\text{Propensity} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^p \beta_i x_i)}},$$

is the most frequently used method to produce propensity scores, where $\beta_0 + \sum_{i=1}^p \beta_i x_i$ is a linear combination of the measured covariates. However, since we are really only interested in the association between covariates and group selection, other predictive methods aside from logistic regression should work (Rosenbaum & Rubin, 1983).

Propensity scores can be used in a variety of ways, including weighting the sample. Typically, the inverse of the scores are used in this case, as we want to emphasize those individuals who are less common in the treatment group. For example, say we were to re-analyze the amount of cigarettes from the weighting sample. Rather than multiplying by the proportion of constituents in the age group, we could calculate each individual's propensity score, taking into account all other measured covariates rather than just age.

We could also apply this method to modeling approaches by adding propensity scores as a covariate adjustment. This would ideally have the effect of controlling for all variables that influence selection into treatment group (Rosenbaum & Rubin, 1983).

Propensity score matching is another popular method to deal with observational data. As one of the initial uses of propensity scores, Judea Pearl even references it as “the most developed and popular strategy for causal analysis in observational studies” (Pearl, 2009). Unfortunately, there is some evidence that propensity score matching actually increases bias, especially when the sample is already fairly balanced. Those interested in exploring why should read Gary King’s “Why Propensity Scores Should Not Be Used in Matching” (2016). However, for the purposes of this thesis, readers should know that this issue is unique to propensity score matching and thus does not carry over to other uses of propensity scores.

The other major limitation with propensity scores in general is that the models that create them are subject to the same limitations that plague the rest of the data and must be specified in advance to be replicable. Improper specification of the propensity model will create biased scores, which will only serve to increase bias in the results (King & Nielsen, in progress).

2.3.6 Instrumental Variables

First described by Philip Wright (1926), *instrumental variables* (IV) are a favored method of economists and many social scientists.

A variable, Z , is said to be an instrument of the treatment X if it satisfies three criteria:

1. Z is associated with X .
2. Z does not affect the outcome, Y , except through X .
3. Z and Y do not share a cause.

Criteria 1 is simple to verify, but the others are not. In fact, it is generally impossible to empirically verify criteria 2 and 3. Researchers must use their own expert knowledge to examine the validity of these assumptions.

Figure 2.4 depicts use of IV, where the dotted line between Z and X indicate some type of active path (causal or not) and L contains all confounders. If there is an active causal path, Z is known as a causal instrument but if it is merely an associational path, Z is a proxy instrument (Hernán & Robins, in progress).

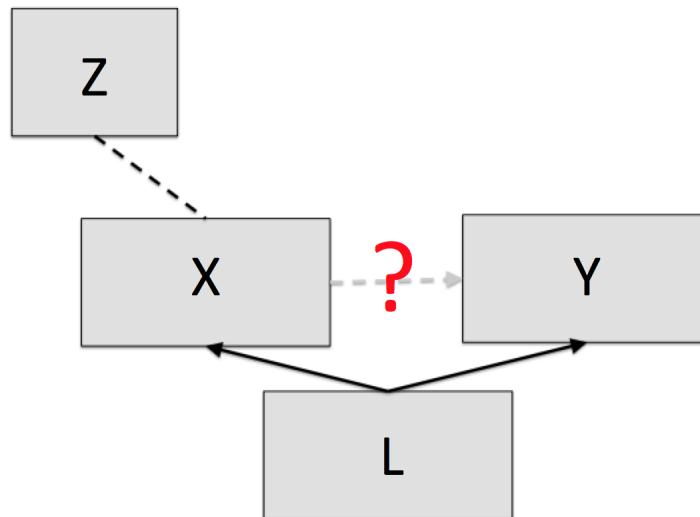


Figure 2.4: Instrumental variable example. An instrumental variable allows researchers to avoid confounding by using an external covariate to predict an untainted expectation of the treatment. While generally not used by epidemiologists and statisticians, it is a favored method of economists and other social scientists.

Now that we have defined an IV, how can we use it? The key to instrumental variables is that they allow us to bypass the confounding effects of L , as we use the instrumental variable Z to predict an untainted expectation of X rather than the true level of treatment.

Say, for instance, that we wanted to determine the effect of counseling on depression in college students. We cannot simply measure depression in students who have and have not received treatment. Students who seek help may be more inclined to benefit

from it, as they have a greater wish to be cured. Instead, we may be able to use students' thriftiness to our advantage by using free counseling as an instrumental variable. It is reasonable to expect that students who can receive free help are more likely to get it than those who have no such option. Assume that the choice to offer it for free is not a consequence of the administration noticing increased depression rates and that schools differ only in the cost of counseling. Thus free counseling and depression are truly independent and it is a good instrumental variable. In a sense, we are using our instrument as a predictor to generate an expected value of treatment that is free from the influence of any confounders or selection bias. We can then compare depression rates in counterfactuals determined by our instrument.

Note that there is an obvious flaw in this method. If Z is not an accurate predictor of X , then the entire analysis may be invalid. In our example, if students are not more likely to get treatment when it is free, then there is nothing between our counterfactuals to compare. The worse our instrument, the weaker the comparison. For this reason, it is common practice to calculate what is known as the *usual instrumental estimand*:

$$\frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[A|Z = 1] - E[A|Z = 0]}$$

Here, we inflate the causal relationship based on the weakness of our instrument. If Z is a perfect predictor of A , then the denominator is 1 and we have an ideal estimation. The worse our instrument is, the closer the denominator is to zero and the greater we inflate our counterfactual comparison (Hernán & Robins, in progress).

Chapter 3

Simulation Study and Applications

Theoretical explanations are necessary aspects of learning, but examples are complementary and equally important, allowing for the description of material in such a way that it is easier to remember. For this reason, the two following examples are included. The first is a simulation study. As the sample is artificially generated, all aspects can be controlled and both counterfactual outcomes are known. This will be useful to explore the challenges of causal inference and illustrate exactly how our various methods work. The second example is from real life data and will be covered in the Application section.

3.1 Simulation Study

3.1.1 Creating A Hypothetical Population

The first step in creating a simulation study is to define a population. This population will consist of 10,000 individuals and will contain information on six variables: sex, smoking status, number of cigarettes per week, number of cups of coffee per week, baseline cancer risk score (without accounting for smoking), and cancer risk score under the conditional assumption that the participant smokes. The fact that both of these risk scores are known is one of the primary advantages of a simulation study. With both counterfactuals, we can calculate a true individual effect measure and compare it to the observed effects of smoking on cancer risk. Thus, we can estimate the accuracy of our model. Table 3.1 describes the variables, as well as a brief explanation of how they were generated. Table 3.2 offers a glimpse of the data. The code used to generate this population can be found in the appendix.

3.1.2 The Search for a Confounder

Observation vs Reality

One of the most difficult aspects of causal inference is that researchers rarely get all the information they need. In fact, lack of observable counterfactuals is a fundamental problem of causal inference (Holland, 1986). This section will examine how this loss

Variable	Description
Sex	Individual's sex, binary Male/Female 65% female
Smoker	Smoking status, binary smoker/nonsmoker 77% of males and 25% of females smoke
Cigarette Base	Number of cigarettes smoked per week, if they were to smoke Males smoke about 20 cigarettes a day, while females smoke on average 7
Coffee	Binary indicator for whether or not they drink coffee 75% of smokers and 15% of non smokers drink coffee in this population
Cancer Baseline Risk	Counterfactual cancer risk score given the person does not smoke Randomly generated from a normal distribution of mean 25 and variance 10
Cancer Risk If Smokes	Counterfactual cancer risk score given the person smokes Generated from function: $cancer_smoke = cancer_base + 0.77cig_base + \epsilon$

Table 3.1: Population characteristics of the simulation study. Values based on true phenomena, exaggerated slightly to allow for more obvious conclusions.

Sex	Smoker	Number of Cigarettes	Coffee	Cancer Baseline Risk	Risk If Smokes
M	1	22	1	16.70	34.11
F	0	5	0	19.05	23.37
F	0	5	0	30.39	34.71
M	0	21	0	20.62	37.26
M	0	28	0	21.97	44.00
F	1	4	1	20.23	23.78

Table 3.2: Sample of true population data in simulation study. One of the primary advantages of a simulation study is that both counterfactuals are known. Thus, we can calculate a true effect for each individual and measure the accuracy at each step in our analysis.

of information can affect uncontrolled results.

Suppose that these data arise from a setting where only one of the counterfactuals is observable. Table 3.3 shows the observable data. Notice that there is only one cancer risk score per participant. Now, rather than directly comparing an individual's potential outcomes, an artificial comparison must be made.

An investigator may first wish to determine if there is a significant difference in cancer risk scores between people who drink coffee and those who do not. One of the most common methods for such a comparison is the two sample t-test. However, a univariate test that does not account for confounders may not produce the expected result. Table 3.4 shows the output of the t-test using the observable data. Although there is a difference in risk score between coffee drinkers and non-drinkers, this is likely a spurious association since coffee drinking played no role in generating cancer risk score. Rather, coffee is correlated with smoking, which does affect cancer risk.

Sex	Smoker	Cigarette Base	Coffee	Observed Risk
M	1	22	1	34.11
F	0	0	0	19.05
F	0	0	0	30.39
M	0	0	0	20.62
M	0	0	0	21.97
F	1	4	1	23.78

Table 3.3: Sample of observable population data in simulation study. Only one counterfactual is visible now. Thus, to get an estimate of the causal effect, we need to use the observed counterfactual of the control group to approximate the unobservable counterfactual of the treatment group.

Stratification

The researchers in the aforementioned study would hopefully investigate their observed relationship further. For instance, they may believe that sex is serving as an effect modifier. After all, females in the study tended to drink less coffee, but were more likely to participate in the study (Table 3.1). The presence of an effect modifier suggests that the artificial counterfactuals are not perfectly exchangeable. In other words, the observed coffee drinking population is not representative of the population as a whole.

To address the problem of differences due to sex, investigators may choose to stratify. In doing so, they will run separate models for males and females. Any differences in the coefficients will support the hypothesis of sex as an effect modifier. Results are shown in Table 3.4. It turns out that sex is an effect modifier with a stronger effect in the male population, but the research team has still not found the true cause of the increased cancer risk.

Stratified	Sex	Estimate (CI)	T-Statistic	P-Value
No		-7.37 (-7.84,-6.90)	-30.72	<0.001
Yes	Male	-7.47 (-8.31,-6.64)	-17.5	<0.001
	Female	-3.00 (-3.55,-2.45)	-10.67	<0.001

Table 3.4: Results from two sample t-test comparing observed cancer risk in coffee drinkers to non-drinkers. Stratifying by sex reveals the presence of an effect modifier, but coffee still appears to cause cancer.

Before moving on to other methods, it may be helpful to recall what the presence of an effect modifier indicates. An effect modifier *is not a cause*, but rather a characteristic of the population across which the causal effect may vary. Sex, by definition, cannot be a cause as we (generally) cannot intervene and change it. Thus there is no counterfactual to compare. This does not mean sex is useless however. Observing that males have a generally higher risk of developing cancer can be a

significant clinical finding. This information may make it easier to identify high-risk individuals, leading to better targeted treatments or outreach programs. It may also improve the causal estimates by restricting counterfactual comparisons to similar people. Yet there are limits to the utility of stratification, primarily its inability to deal with continuous and high dimensional settings. Other, more flexible methods are also needed.

Regression Model

Regression models are some of the most ubiquitous methods in all of statistics. They are also integral tools of causal inference, particularly for identifying confounders. To illustrate their use, we will continue with our coffee-cancer example.

Suppose that investigators decide to try a linear regression model. They may create an initial model of the form,

$$\text{cancer risk} = \beta_0 + \beta_1 \text{coffee} + \epsilon,$$

Where β_0 is the baseline risk for cancer without accounting for any other variables, β_1 is the additional cancer risk associated with smoking, and ϵ is the error term. When this model is fit to the data, coffee appears to be a significant, albeit modest, predictor (Model 1 in Table 3.5). Even after incorporating sex into the model, coffee remains significant (Model 2 in Table 3.5, these results are consistent with 3.4). If those are the only variables that the researchers decide to look at, they are likely to reach a false conclusion.

Luckily for the researchers, they did not stop at Model 2 with only coffee and sex. Due to their expert knowledge of the field, they had reason to believe that smoking may also be a factor. Thus, they included smoking status into their model (Model 3 in Table 3.5).

At this point, coffee drops out as a significant predictor of cancer risk. After conditioning on smoking status, there is no active path linking coffee and cancer. Yet, we have not proved that coffee is not a cause of cancer. It is plausible that coffee leads to smoking, which then causes cancer. From the data we have so far, we cannot rule this out. In fact, even if we had data for the entire population, with both the smoking and non-smoking counterfactuals for the cancer risk score visible, it would be impossible to say whether coffee was an ancestral cause. It is only by knowing the data generation method, $\text{cancer risk} = \text{baseline} + 0.77 * \text{number of cigs} + \epsilon$ and (more importantly) that smoking status determined the propensity to drink coffee, that we can determine that smoking is a common cause (3.1). This is one of the main reasons why causal inference is so difficult and, in some cases, impossible.

It is interesting to note that even if we cannot fully rule out coffee as a cause of cancer, these models can have some clinical significance. Knowing that smoking is further down the causal chain, closer to the true cause of cancer, may allow for more targeted intervention in the future. For instance, researchers may specifically recruit participants from coffee shops, knowing that they are likely to smoke and be at higher risk for cancer. Even if coffee is not a cause of cancer, it can be a reliable predictor.

	<i>Dependent variable:</i>		
	Observed Cancer Risk		
	Model 1	Model 2	Model 3
Constant	27.1 (0.2)*	25.0 (0.2)*	24.1 (0.1)*
Coffee Drinker	7.3 (0.2)*	4.7 (0.2)*	0.04(0.3)
Sex (Male)		9.2 (0.2)*	5.7 (0.3)*
Smoker			9.3 (0.3)*
R ²	0.088	0.203	0.280
Residual SE	11.7	10.9	10.4
F Statistic	965*	1,277*	1,294*
<i>Note:</i>			*p<0.01

Table 3.5: Initial models to describe cancer risk score. Notice that coffee drinking is initially significant, but is not after smoking status is added to the model (marked in red). This shows that smoking is a mediator or common cause in the relationship between coffee consumption and cancer.

Dose Effect

Notice that even though Model 3 successfully identifies smoking as a predictor and eliminates coffee, sex is still a significant predictor in the model. Yet sex did not play a role in generating the cancer risk score (Table 3.1). In fact, we should be able to exactly reproduce the generating model if we incorporate the correct variables. This illustrates another important topic in causal inference, the dose effect. Not all smokers are created equal. Imagine two smokers, one who smokes a cigarette at a bar and another who goes through a pack a day. Should these people be expected to have a similar increase in cancer risk? Obviously not.¹

The smoking cancer risk score in this population was generated based on the number of cigarettes the person smoked per day, but Model 3 in Table 3.5 used a dichotomous smoking indicator. This leads to sex retaining its significance, as males tended to smoke a greater quantity of cigarettes than females. If number of cigarettes is used in place of smoking status, sex should drop out as a predictor. Model 4 from Table 3.6 shows this to be true. Furthermore, if number of cigarettes is used as the sole predictor of cancer risk (Model 5, Table 3.6), the exact underlying model can be recovered.

3.2 Applications

There are many examples of using observational data used to solve practical questions. Some, like the effect of smoking on cancer, have such a preponderance of evidence and support that their effects are indisputable. Others, like bone marrow replacement to treat leukemia have proven to be dangerously misleading. Table 3.7 shows some results from observational studies that were validated or contradicted by subsequent randomized controlled trials (Lauer, 2011).

3.2.1 Clearly Wrong: Bone Marrow Transplant

Cancer is one of the deadliest and most ubiquitous diseases known to man. The National Cancer Institute estimates that nearly 40% of people will be diagnosed with cancer at some point during their lifetime and it is listed as the second leading cause of death, behind only heart disease (Jemal, 2017). Breast cancer is the most common form of cancer, with an estimated 250,000 new cases per year (Cancer Facts and Figures, 2018). With such a large public health burden, it is understandable that cancer is the most heavily researched field of medicine, accounting for \$5.6 billion (16%) of total NIH funding (2017).

In the 1980s, bone marrow transplants to treat breast cancer were widely prescribed. Based on Nobel Prize-winning research on its effectiveness in another type of cancer, leukemia, and a couple of small observational trials, nearly every academic center in the country had an established program (Thomas et al., 1975). In fact it is estimated

¹We did choose to dichotomize smoking earlier, but that was to simulate a real environment where smoking is often encoding as such. More information is always preferable.

	<i>Dependent variable:</i>	
	Observed Cancer Risk	
	Model 4	Model 5
Constant	25.0 (0.1)*	25.0 (0.1)*
Sex (M)	-0.34 (0.3)	
Cigarette	0.81 (0.01)*	0.79 (0.01)*
R ²	0.33	0.33
Residual Std. Error	9.99	9.99
F Statistic	2,487*	4,973*
<i>Note:</i>	*p<0.01	

Table 3.6: Models after accounting for dose effect. Cancer risk was initially generated by a linear combination of the patient's baseline cancer risk and the number of cigarettes they smoke per week. Thus, although sex was significant when smoking was measured as a dichotomous indicator, it was not when the more specific measurement was used. Moreover, once the regression model was left with only the number of cigarettes variable, the generating model was recovered, with the intercept serving as the average baseline cancer risk score in the population.

Validated	Contradicted
Lower blood pressure with drugs Lower LDL cholesterol with statins Aspirin to prevent MI and stroke Beta-blockers and ACE inhibitors for CHF Mammography for breast cancer CT for lung cancer	Vitamins to prevent cancer/ CVD (failed) Anti-arrhythmic drugs (higher death rates) Back surgery, kyphoplasty (little benefit) Aggressive glucose reduction to prevent MI Stents <i>after</i> MI Bone marrow transplant for breast cancer

Table 3.7: Hypotheses from observational tests after further study in RCT. Some have proved true, others have not. Example taken from Dr. Michael Lauer’s presentation at the International Society for Pharmacoeconomics and Outcomes Research.

that between the years 1985-1988, more than 30,000 women received this treatment. However, despite its popularity, there had been no large scale randomized controlled studies to confirm the effectiveness of bone marrow transplants on solid tumors. Its effectiveness was largely just assumed to carry over (Adkins et al., 1999).

Finally, at an American Society of Clinical Oncology meeting in 1999, 4 out of 5 newly published abstracts of true RCTs showed little to no benefit in bone marrow transplants to treat breast cancer (Rettig, Jacobson, Farquhar, & Aubry, 2007). Soon after, Stadtmauer and colleagues (2000) published the results of a large scale randomized controlled trial in the *New England Journal of Medicine*. They also found no clinical benefit.

The randomized trials were clear about their findings: the treatment simply did not work. As it was also very expensive, bone marrow transplantation in breast cancer patients was quickly phased out of practice.² However, not all implications of RCTs are so clear. The practice of hormone replacement therapy for post menopausal women to treat heart disease is a decidedly more complex example.

3.2.2 Causality is Difficult: Hormone Replacement

Of all the results based on observational data, perhaps no story is as fascinating as the use of hormone replacement therapy (HRT) to treat heart disease. This series of events illustrates how difficult it can be to draw causal inference and the unanticipated effects these results can have. The story is messy and still not resolved but is important to understand the potential challenges and consequences of causal inference.

It is a well known fact that as women grow older, the hormones in their body begin to change. For a time, it was thought that these hormonal shifts were partially responsible for a variety of heart and other health problems. These theories were backed up by results from the Nurses’ Health Study (NHS), a large observational trial that collected data over time from thousands of registered nurses in the New England

²The economics of bone marrow transplant are also fascinating, as the procedure was very expensive. Readers interested in health economics and insurance policy are encouraged to read Rettig’s 2007 book, *False Hope: Bone Marrow Transplantation for Breast Cancer*.

area (Stampfer et al., 1991, 1985). At one point, it is estimated that nearly 60 percent of post-menopausal women were receiving HRT and it was seen as the standard of care (Herrington & Howard, 2003).

Unfortunately, later randomized experiments from the Womens' Health Initiative (WHI) suggested that HRT was ineffective. In fact, not only did HRT fail to help, it actually increased patients' risk for developing certain heart conditions and cancer (Stefanick et al., 2003). These findings led to tremendous public outcry. Shortly after the initial report, the UK's MRC estimated that HRT had caused 20,000 deaths from breast cancer over the past ten years in the UK alone (Brower, 2003). Bruno Muller-Oerlinghausen, chairman of the German Commission on the Safety of Medicines called the thousands of breast cancer deaths from HRT "a national and international tragedy" and stated in the *UK Sunday Herald* that "more women have probably died from the hormone therapy than damaged children were born in the wake of the thalidomide scandal" (Burgermeister, 2003).

Ultimately, further research proved the WHI results to be misleading. It turned out that the populations in the two studies were not exactly the same. Women in the NHS were generally younger and more fit than those in the WHI. Moreover, HRT does seem to have an effect if given in small doses immediately following menopause. This phenomenon, known as the timing effect, helps to explain why the initial cohort study showed positive results, while the later RCT did not. People in the NHS just happened to be in the ideal target population (Langer, Manson, & Allison, 2012). Interestingly, the initial report by the WHI actually stated that there were some methodological differences between the studies and warned against rash comparisons. They expected their work to be part of a larger body of literature on the subject and hoped to inspire a critical reflection on the use of HRT. But observers did not heed their warnings (Stefanick et al., 2003).

Today, there are better treatment options and HRT is not commonly prescribed (Lobo, 2017). However, there are many lessons that can be learned by examining its story. The first and most obvious lesson, is to always be careful and observant when analyzing an experiment. No study is perfect and it is important to critically evaluate its flaws before relying on its conclusions in practice. Second, do not discount results gained from observational studies or place undue faith in a randomized trial. Experiments can also be flawed and misleading.

3.2.3 Real Data Example

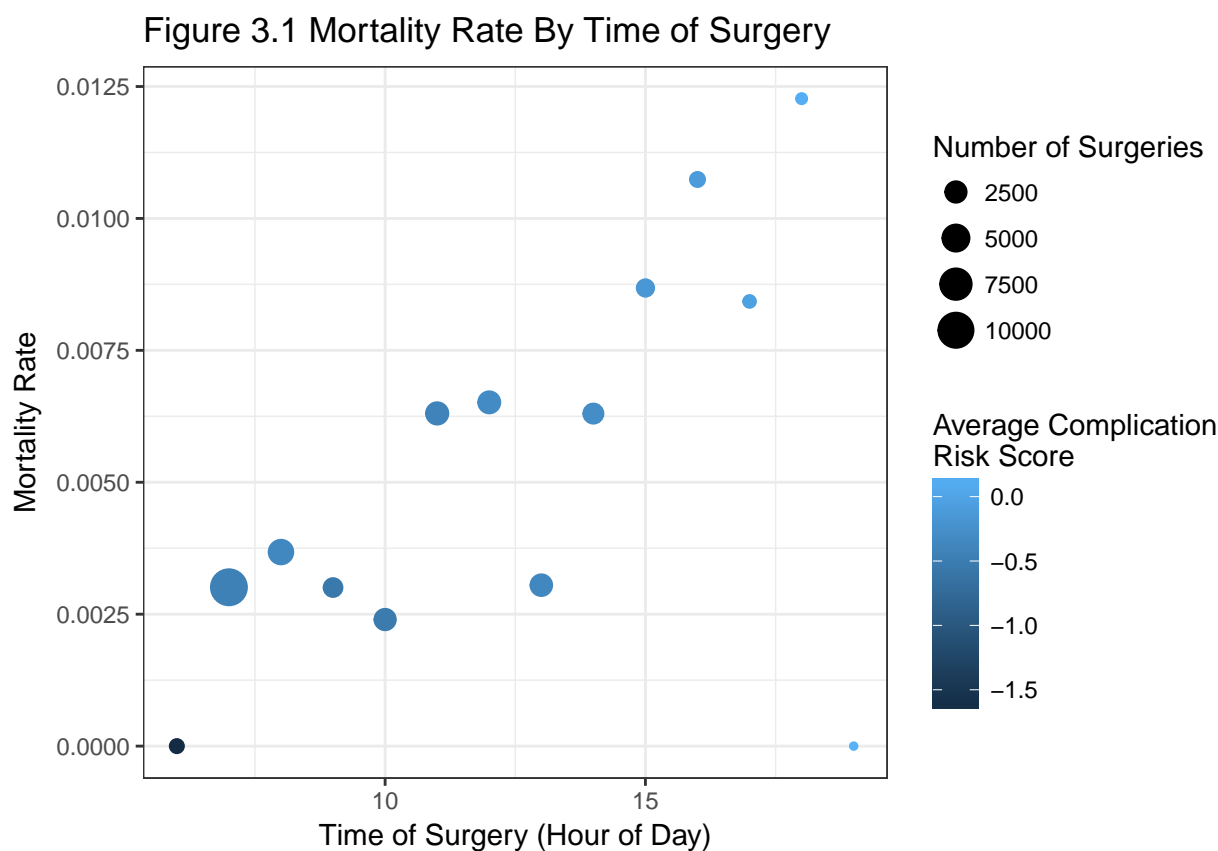
So far we have examined simulated examples and descriptions of real causal studies. While these are important steps to understand causal inference, it may also be useful to carry out the methods described in this thesis.

Data for this example was taken from the American Statistical Association's Teaching Statistics in the Health Science's Resources Portal. The dataset was uploaded in 2016 by Amy Nowacki and documents 32,001 elective general surgery patients. Investigators in this study hoped to determine if the time of day of a procedure had an effect on outcomes, specifically complication and mortality rate (Nowacki, 2016). This example will strive to disprove the hypothesis of timing having an effect on surgery

outcomes with a methodology similar to the Searching for Confounders section of the simulation study.

Data exploration

The first step in an analysis is to determine if the hypothesis is reasonable. In this case, we should explore whether there is an actual difference in mortality over time. Figure 3.1 is a plot of time vs mortality rate, with the continuous time variable broken up into hour long, discrete segments to allow mortality rate to be calculated. It turns out that mortality does increase later in the day. Some of this change may be due to a slight rise in pre-surgery mortality risk scores, as surgeons may simply decide to take on tougher cases in the afternoon.



Model Building

Say we did not yet know that mortality rate rose in the afternoon. An initial model with only hour of the day as a predictor would be significant (see Table 3.8). However, as soon as some basic patient characteristics are included, hour loses its significance. Thus, we can see that this variation can be explained by differences in the population.

	<i>Dependent variable:</i>			
	30-Day Mortality		Complication Rate	
	Hour Only	W/ Patient Info	Hour Only	W/ Patient Info
Constant	−5.7 (0.14)*	−7.9 (0.51)*	−1.9 (0.09)*	−2.4 (0.08)*
Age		0.028 (0.01)*		0.012 (0.00)*
Afternoon	0.6 (0.10)*	0.2 (0.21)	0.1 (0.04)*	−0.00 (0.04)
Complication Risk		0.6 (0.04)*		0.3 (0.02)*
Mortality Risk		0.3 (0.10)*		0.2 (0.02)*
Log Likelihood	−883	−558	−12,555	−11,912
Akaike Inf. Crit.	1,770	1,126	25,114	23,834

Note:

***p<0.01**

Table 3.8: Initial models from real data: hour only and hour with patient characteristics. Once patient characteristics were added in, time of day lost significance in both the mortality and complication rate models.

Propensity Scores

There are a lot of variables included in the surgery dataset that were not included in the earlier models. In situations like this, propensity methods are common.

Propensity scores are generally designed to deal with two levels of treatment. For the purposes of this example, we will temporarily dichotomize the *hour* variable to morning and afternoon. Then, we will use all variables except for the outcome variables of interest (30-day mortality and 30-day complication rate) as coefficients in a logistic regression model to predict the probability a patient would have their surgery in the afternoon. This generates our propensity score (code shown below).

```
# A better analysis would include an examination of the missing data
# To avoid selection bias. We will just drop them.
surgeryP <- surgery[complete.cases(surgery),]

# Propensity scores are generated from a logistic regression model
# predicting probability of receiving treatment
mProp <- glm(afternoon ~ . - mort30 - complication - hour,
             family = "binomial", data = surgeryP)

surgeryP <- surgeryP %>% mutate(pscore = mProp$fitted.values)
```

Adjustment

Once we have a propensity score, there are several ways to use it. The first is to use it as a coefficient in a linear regression model, along side the treatment and any other variables not included in the propensity-generating function. In this example, all non-treatment or outcome variables were included in the propensity scores, so there will be only two variables in the model.

3.9 shows the results of models with propensity scores. These models disprove the timing hypothesis similarly to those that included all patient information.

This method of propensity score adjustment is a common, effective and user-friendly way to identify a causal effect, even in the presence of confounders in observational data.

3.2.4 Conclusion

Causal inference in practice is very difficult. Real world data is often lacking in detail that may not be obvious to those without expert knowledge. If these limitations violate our key assumptions it is likely that an incorrect conclusion will be drawn. Researchers should remain cautious about relying on these assumptions, but, as in the case of HRT, should not immediately assume their failure in favor of a randomized trial. Each new study adds some piece of information. We need to be informed enough to make use of it all.

	<i>Dependent variable:</i>					
	30-Day Mortality			Complication Rate		
	Hour Only	W/ Pt Info	W/ Propensity	Hour Only	W/ Pt Info	W/ Propensity
Constant	-5.6 (0.1)*	-7.9 (0.5)*	-9.7 (0.4)*	-1.9 (0.0)*	-2.4 (0.1)*	-3.0 (0.1)*
Age		0.02 (0.01)*			0.01 (0.00)*	
Afternoon	0.6 (0.2)*	0.2 (0.2)	0.3 (0.2)	0.1 (0.0)*	0.0 (0.0)	0.0 (0.0)
Complication Risk		0.6 (0.0)*			0.3 (0.0)*	
Mortality Risk		0.3 (0.1)*			0.2 (0.0)*	
Propensity Score			11.9 (0.9)*			3.6 (0.2)*
Log Likelihood	-883	-558	-682	-12,555	-11,912	-10,924
Akaike Inf. Crit.	1,770	1,126	1,370	25,114	23,834	21,854
<i>Note:</i>						*p<0.01

Table 3.9: Propensity score as coefficient adjustments. Models including propensity scores perform similarly to those controlling for each covariate separately.

Chapter 4

Conclusion

A primary goal of science is to move beyond questions of associations to a better understanding of cause and effect. While this event may be associated with this effect, did it cause it? While conceptually straightforward and, perhaps, intuitive, the statistical approach to causal inference is a broad and dynamic field that continues to rapidly evolve.

The goal of this thesis is to provide readers with a fundamental understanding of causal inference — to help readers contextualize claims of causality and evaluate their validity. The information in this work can be applied to a variety of disciplines, even for those who do not plan to independently design an experiment or run a regression model. However, the need for a basic understanding of causal inference is, perhaps, universal, and a conceptual understanding of counterfactuals may be very useful.

For readers who follow the fields of the health, behavioral and social sciences and review the scientific literature, understanding the concepts of counterfactuals will help to contextualize research findings and analyze the quality of the study. The use of DAGs to diagram the relationship between potential causes and effects may help to identify potential confounders that may undermine the conclusions. To achieve these skills, a strong conceptual understanding of the topics discussed here may be helpful.

For those readers who wish to study causal inference in more depth as they design experiments or analyse data for causality, this thesis can serve as a starting point, providing a basic foundation from which an exploration of more advance topics and methods can be launched. Much of the material this in thesis was based on core components of early graduate study in causal inference. In particular, Hernán and Robins', "Causal Inference", is widely regarded as a comprehensive introductory text and serves as the basis for Harvard's *Introduction to Epidemiology* course, among others. Reading this book will reinforce many of the concepts discussed here, but with a more nuanced and technical focus. Those interested in analytical methods may also wish to read Roy's paper on the advance FLAME matching algorithm (2017) for both its intriguing use of machine learning in causal inference and comprehensive discussion of other matching methods. In addition, King's discussion of the dangers of propensity score matching (2016) is important for its overview of the limitations of both propensity scores and matching techniques. Other topics of interest may be multileveled propensity scores, different experimental designs (such as clustered

randomized sampling), and popular methods in healthcare such as survival analysis.

An understanding of causal inference is a core component of the scientific method and, in many ways, an essential step on our journey to understand the world around us.

Appendix

4.1 Definitions From Various Texts

Some texts define the terms confounding, sampling and selection differently. As students may see these terms used in different contexts, we will include a brief discussion here (for complete definitions see 4.1). Books included are those that are listed as the primary texts for Amherst College's statistics courses for the 2018 spring semester. Also included are Hernán and Robins' "Causal Inference", the primary text for Amherst College's "Research Methods" course in the Psychology Department and the Wikipedia entries for confounding, sampling bias and selection bias. Other epidemiology texts were not included due to space limitations and their general overlap with Hernán.

Confounding: The introductory books have a definition of confounding that is close to Hernán's, but do not differentiate between common causes and mediators. Thus, not every confounder will be non-causal. In fact, De Veaux explicitly allows for this ambiguity by defining a separate term, lurking variables, that fits Hernán's definition.

Sampling: The introductory texts briefly covered methods to sample, but this is a topic that is generally not covered in books devoted to causal inference.

Selection: This is the section with the most variation. The psychology text conflates this with confounding, while it seems to generally not be considered important to include in introductory texts. The Wikipedia page seems to combine both sampling bias and confounding in its definition.

4.2 Code Used to Create Simulation Study Population

```
# Note: Code will be shown in appendix once simulation is okay

set.seed(777)

##### This is where we will create our simulated population

# There will be 6 variables:
```

```

# sex:           Sex of the individual
#               (Binary M/F for simplification purposes)
# smoker:       Smoking status
#               (Binary 1/0 for smokers and nonsmokers, respectively)
# cig_base:     Number of cigarettes smoked per week if person smoked
# coffee:       Number of cups of coffee the person drinks per week
# cancer_base:  Counterfactual cancer risk score if person did not smoke
# cancer_smoke: Counterfactual cancer risk score if person did smoke

##### Here we will define our constants

# Number of people
n <- 10000
# Percent Female in Study
percentF <- 0.65
# Smoking Statistics
## Mean number of cigarettes per week by sex
cigM <- 20
cigF <- 7
## Proportion of smokers by sex
propCigM <- 0.77
propCigF <- 0.25
# Proportion of Coffee Drinkers by smoking status
coffeeS <- 0.75
coffeeNS <- 0.15
# Variable coefficient for smoking in model
beta_cig <- 0.77

##### Now we will generate the population

pop <- data_frame(sex = ifelse(runif(n) > percentF, "M", "F"),
  # Baseline cancer risk
  cancer_base = rnorm(n, 25, 10),
  # Number of Cigarettes if smoker
  # Different rates for male/female
  # Also, dependent on patient age
  cig_base = ifelse(sex == "F", rpois(n, cigF),
    rpois(n, cigM))
) %>%
mutate(
  # Cancer Risk with covariates

```

```
# Linear combination of covariates to get a "nurture" risk
# Also include a normal error term
err_var = var((cig_base * beta_cig) / 7),
cancer_smoke = cancer_base + (cig_base * beta_cig) + rnorm(1, 0, err_var),

# Determine which counterfactual is observed
smoker = ifelse(runif(n) < ifelse(sex == "M", propCigM, propCigF), 1, 0),

# Generate coffee drinking status
coffee = ifelse(runif(n) < ifelse(smoker == 1, coffeeS, coffeeNS), 1, 0)
)

# Reorder columns to make sense
pop <- pop[,c(1,6,3,7,2,5)]
```

Level	Title (Author)	Confounding Variable Definition	Sampling	Selection
Masters in Epidemiology	Causal Inference (Hernán and Robins)	There exists a common cause creating a non-causal, active path between the treatment and response	Briefly	Yes
Intro (No Modeling)	Seeing Through Statistics (Utts)	Related to explanatory (x) variable and affects the response (y) variable	Yes	No
Intro (With Modeling)	Stats: Data and Models (De Veaux)	Associated in a non causal way with a factor and affects the response Distinctly defines lurking variables as associated with both x and y that makes it appear x causes y	Yes	No
Intermediate	Stat2 (Cannon)	No Mention	No	No
Math Stats	Mathematical Statistics and Data Analysis (Rice)	Variable that causes an imbalance between treatment and control groups	Not	No
Psychology Research Methods	Research Methods in Psychology (Morling)	Alternative, non-causal explanation to observed effect due to problematic selection	Yes	No
Wikipedia	Confounders	A variable that influences both the dependent and independent variables causing a spurious association		
Wikipedia	Selection Bias	Bias introduced by the selection of individuals in such a way that proper randomization is not achieved thereby ensuring an unrepresentative sample		
Wikipedia	Sampling Bias	Bias from a sample collected in such a way that some members of the intended population are less likely to be included than others		

Table 4.1: Definitions of confounding and discussion of sampling or epidemiological selection bias from various texts.

References

- Adkins, D., Brown, R., Trinkaus, K., Maziarz, R., Luedke, S., & Freytes, C. (1999). Outcomes of high-dose chemotherapy and autologous stem-cell transplantation in state IIIB inflammatory breast cancer. *Journal of Clinical Oncology*, 17(7).
- An, W., & Ding, Y. (2017). The landscape of causal inference: Perspective from citation network analysis. *The American Statistician*, 70(1), 41–55.
- Beery, A., & Zucker, I. (2011). Sex bias in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews*, 35(3), 565–572.
- Brower, V. (2003). A second chance for hormone replacement therapy. *EMBO Reports*, 12(4), 1112–1115.
- Burgermeister, J. (2003). Head of German Medicines Body likens HRT to thalidomide. *British Medical Journal*, 327(7418), 767.
- Fisher, R. (1920). Notes on the history of correlation. *Biometrika*, 13(2), 25–45.
- Frieden, T. (2017). Evidence for health decision making - beyond randomized controlled trials. *New England Journal of Medicine*, 377, 465–475.
- Guyatt, G. (1991). Evidence-based medicine. *Annals Internal Medicine*, 14(2), A–16.
- Guyatt, G. (2010). GRADE guidelines: 4. rating the quality of evidence - study limitations (risk of bias). *Journal of Clinical Epidemiology*, 407–415.
- Hernán, M. A., & Robins, J. M. (in progress). *Causal inference*. Chapman & Hall/CRC.
- Herrington, D., & Howard, T. (2003). From presumed benefit to potential harm hormone therapy and heart disease. *New England Journal of Medicine*, 349(6), 519–520.
- Hill, S. A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American*

- Statistical Association*, 81(396), 945–960.
- Hume, D. (1740). *A treatise on human nature*.
- Jemal, A. (2017). Annual report to the nation on the status of cancer 1975-2014. *Journal of the National Cancer Institute*, 109(9).
- King, G., & Nielsen, R. (in progress). Why propensity scores should not be used in matching.
- Langer, R., Manson, J., & Allison, M. (2012). Have we come full circle or moved forward the Womens Health Initiative 10 years on. *Climacteric*, 15(3), 206–212.
- Lauer, M. (2011). Randomized trials vs. observational studies. International Society for Pharmacoeconomics; Outcomes Research.
- Lilford, R. J., & Jackson, J. (1995). Equipose and the ethics of randomization. *Journal of the Royal Society of Medicine*, 88(10), 552–559.
- Little, R., & Rubin, D. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, 21, 121–145.
- Lobo, R. (2017). Hormone replacement therapy current thinking. *Nature Reviews Endocrinology*, 13, 220–231.
- Michelle Irving, N. C., Ranmalee Eramudugolia, & Anstey, K. (2017). A critical review of grading systems: Implications for public health policy. *Evaluation and the Health Professions*, 40, 244–262.
- Nowacki, A. S. (2016). Surgery timing dataset. *TSHS Resources Portal*.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan; Kaufman.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.
- Peng, R., Dominici, F., & Zeger, S. (2006). Reproducible epidemiological research. *American Journal of Epidemiology*, 163(9), 783–789.
- Reiter, J. (2000). Using statistics to determine causal relationships. *The American Mathematical Monthly*, 107(1), 24–32.
- Rettig, R., Jacobson, P., Farquhar, C., & Aubry, W. (2007). *False hope: Bone marrow transplantation for breast cancer*. New York: Oxford University Press.
- Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Roy, S., Rudin, C., Volfovsky, A., & Wang, T. (2018). FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference. *ArXiv E-Prints*.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonran-

- domized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.
- Russell, S. J. (2011). A.M turning award 2011. Retrieved March 21, 2018, from https://amturing.acm.org/award_winners/pearl_2658896.cfm
- Sackett, D. (1981). How to read clinical journals: I. why to read them and how to start reading them critically. *Can Med Assoc J*, 124(5), 555–558.
- Smith, G. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomized controlled trials. *British Medical Journal*, 327, 1459.
- Spirtes, P. (1996). *Using d-separation to calculate zero partial correlations in linear models with correlated errors*. Carnegie Mellon University.
- Spirtes, P. (2017, july). Causal & statistical reasoning. Carnegie Mellon University Open Learning Initiative.
- Stampfer, M., Colditz, G., Willett, W., Manson, J., Rosner, B., Speizer, F., & Hennekens, C. (1991). Postmenopausal estrogen therapy and cardiovascular disease ten year followup from the Nurses Health Study. *New England Journal of Medicine*, 325, 756–762.
- Stampfer, M., Willett, W., Colditz, G., Rosner, B., Speizer, F., & Hennekens, C. (1985). A prospective study of postmenopausal estrogen therapy and coronary heart disease. *New England Journal of Medicine*, 313, 1044–1049.
- Stefanick, M., Cochrane, B., Hsia, J., Barad, D., Liu, J., & Johnson, S. (2003). The Womens Health Initiative postmenopausal hormone trials overview and baseline characteristics of participants. *Ann Epidemiol*, 13, S78–S86.
- Suppes, P. (1970). A probabilistic theory of causality.
- Thomas, E. D., Storb, R., Clift, R., Fefer, A., Johnson, F., Neiman, P., . . . Buckner, C. (1975). Bone-marrow transplantation. *New England Journal of Medicine*, 292(16), 832–843.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 10(7), 557–585.