Observational Studies 2 (2017) 28-38 Submitted 8/17; Published 8/17

Book review of "Observation and Experiment: An Introduction to Causal Inference" by Paul R. Rosenbaum

Dylan S. Small **dsmall@wharton.upenn.edu**
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA, U.S.A.

The economist Paul Samuelson said, "My belief is that nothing that can be expressed by mathematics cannot be expressed by careful use of literary words." Paul Rosenbaum brings this perspective to causal inference in his new book *Observation and Experiment: An Introduction to Causal Inference* (Harvard University Press, 2017). The book is a luminous presentation of concepts and strategies for causal inference with a minimum of technical material. An example of how Rosenbaum explains causal inference in a literary way is his use of a passage from Robert Frost's poem "The Road Not Taken" to illuminate how causal questions involve comparing potential outcomes under two or more treatments where we can only see one potential outcome:

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth; (Frost (1916))

"Frost creates the mood attending a decision, one whose full consequences we cannot see or anticipate. 'Knowing how way leads on to way,' we will not see the road not taken. So it was for Frost in a yellow wood...so it is for a patient at risk of death in the ProCESS trial [a randomized trial comparing two treatments for septic shock], so it is in every causal question."

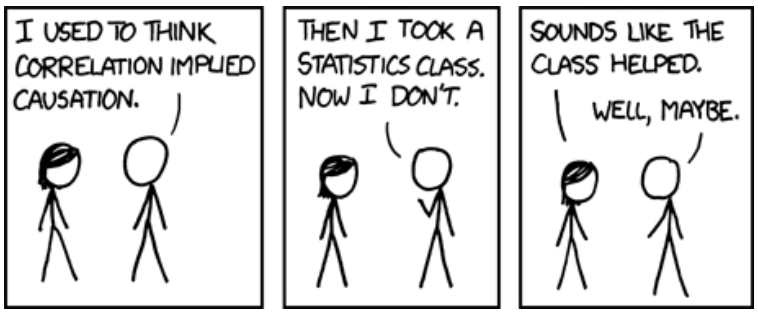
In reverse order of its title, *Observation and Experiment* starts with an account of causal inference from randomized experiments and then moves to observational studies. The randomized experiment is a powerful tool for causal inference – it provides an automatic way to infer the causal effect of a treatment without understanding why different people have different preferences for treatments. It does this by suppressing the role of preferences in choosing treatments – people cede control of their choice to a random coin flip. But in many settings, people refuse to cede control or it would be unethical to try to force them to cede control. We cannot force some people to smoke cigarettes and others not to. "The central problem in an observational study," Rosenbaum says, "– the problem that defines the distinction between a randomized experiment and an observational study is that treatments are not assigned at random...In the US in 2016, the poor are far more likely than the rich to smoke cigarettes, as the foolish are more likely to smoke than the wise. If

poverty and foolish behavior have consequences for health besides increased smoking, an investigator will need to take care and exert effort to isolate the effects actually caused by smoking."

To make causal inferences from observational data, we must confront that different people have different preferences for treatments. Rosenbaum presents strategies and considerations for confronting this problem. One strategy is to look for circumstances which resemble a randomized experiment in which preferences did not play a major role in determining treatment but instead "a process that is haphazard, senseless, without aim or ambition, equitable, symmetrical" – a natural experiment. Does growing up in a poor neighborhood make a child earn less as an adult? Different parents have different preferences and means for where to live, but in Toronto, there was a haphazard element among families applying for public housing – families on the waiting list were assigned to the next available residence, sending families to public housing projects in varied neighborhoods of the city. Oreopoulus (2003) used the waiting list as a natural experiment to study the effect of growing up in a poor vs. not poor neighborhood on adult earnings. Rosenbaum points out that while it is often assumed that waiting lists resemble randomized experiments and create natural experiments, this needs careful case by case consideration. Similar to the Toronto public housing study, natural experiments have been constructed based on the Gautreaux program which sought to assist black Chicago public housing residents living in heavily segregated areas to move into more integrated areas where families on a waiting list were offered units in different areas supposedly on the basis of their rank order on the waiting list without regard to expressed preferences (Rosenbaum, 1995). However, measured family characteristics were associated with neighborhood assignment, e.g., families owning a car were more likely to be assigned to neighborhoods with a higher percentage of whites (Votruba and Kling, 2009). If these families differ in terms of measured covariates like car ownership, then how can we be confident that within strata of the measured covariates, they are the same on unmeasured covariates as they would be in a randomized experiment? In contrast, in the Toronto public housing study, there was no correlation between measured family characteristics and type of neighborhood the family was placed into (Oreopoulus, 2003).

Natural experiments can be constructed, not just "found" by isolating aspects of the data in which a natural experiment occurred. Rosenbaum says, "Over the course of a life, much may be predictable, perhaps rationally planned, perhaps pointlessly habitual, perhaps neurotically determined, perhaps corralled by social convention or politics that punish the smallest deviation. Still, in such a life, there may be brief moments when a fateful choice between two very different paths is decided by little more than luck. The proverb says, 'For want of a nail, the horseshoe was lost, then the rider, the battle and the kingdom. Isolation refers to sifting a large collection of data to collect these rare, brief moments in which chance plays a decisive role in shaping a life." Zubizarreta, Small and Rosenbaum (2014) used isolation to study the effect of number of children a mother has on a mother's career. Many decisions about having children are planned, but when a mother is having a child, chance occasionally intervenes to produce twins rather than a single child.

The ideal observational study is a natural experiment that resembles randomized assignment, perhaps a lottery as in Imbens, Rubin and Sacerdote's (2001) study on the effects on work, savings and consumption of being handed a large pile of cash by com-



Reprinted from xkcd.com

paring winners and losers of a state-run lottery in Massachusetts. However, such lottery natural experiments are not available for many causal questions. Rosenbaum emphasizes that observational studies can provide useful information when supplemented with tools and strategies such as sensitivity analysis, elaboration of a casual theory and quasi-experimental devices. Elementary statistics students are taught "correlation does not imply causation." The cartoon above captures the ambiguity in whether learning this alone is an advance.

Observational studies that indicated that smoking caused lung cancer played a critical role in reducing smoking and are thought today to have come to the correct conclusion. Dismissing all observational studies on smoking and lung cancer because they only show correlations would have significantly harmed public health. Yet, sometimes correlations are far from implying causation. How can we distinguish between when observational evidence suggests causation and when it is less persuasive. One tool is sensitivity analysis, which asks, "How far would we have to depart from randomized treatment assignment to alter the practical or quantitative conclusions of an observational study?" The first sensitivity analysis for an observational study was carried out by Cornfield et al. (1959) for appraising evidence that smoking causes lung cancer. Observational studies had found strong correlations between smoking and lung cancer, for example Dorn (1959) found that smokers had 9 times the risk of dying from lung cancer as nonsmokers. R.A. Fisher (1958) suggested these results might not show smoking causes lung cancer but could be explained by a genetic variant which makes a person more likely to smoke and more likely to contract lung cancer. Cornfield et al. showed that, ignoring sampling variability, in order for the results of Dorn (1959)'s study to not indicate that smoking causes lung cancer and to be purely explained by an unmeasured confounder, the genetic variant would have to be at least nine times as likely among smokers as nonsmokers. The sensitivity analysis replaces the statement "correlation does not imply with causation," with the more useful statement "in order for this correlation to not imply causation, the bias in treatment assignment must exceed a particular magnitude." A genetic variant that is nine times as likely among smokers as nonsmokers might be judged to be unlikely and this would strengthen belief that smoking causes lung cancer. Rosenbaum has developed a sensitivity analysis model that extends Cornfield et al.'s sensitivity analysis to account for sampling variability, observed covariates and different types of outcomes; the model was described in technical detail in Chaper 4 of his book *Observational Studies* (Rosenbaum, 2002) and he describes the model in a nontechnical way in Chapter 9 of *Observation and Experiment*.

When there is a true treatment effect, we would like to be able to report that an observational study is insensitive to a small or moderate amount of bias (i.e, we would like to still have evidence of a treatment effect even allowing for such a bias); this is analogous to how in a randomized trial, we would like to have high power for detecting a true treatment effect. In Chapter 10 of *Observation and Experiment*, Rosenbaum presents strategies for designing observational studies that can make them more insensitive to bias when there is a true treatment effect. One example is that discovering effect modification increases insensitivity to bias. For instance, Hsu, Small and Rosenbaum (2013) found in a study of the effect of a village level intervention against malaria, there was a strong correlation between the intervention and reductions in malaria among young children and a much smaller correlation among adults. In order to explain this finding as resulting purely from an unmeasured village covariate rather than any causal effect of the intervention, it would require an unmeasured village level covariate with a bigger effect than if there had been a moderate correlation between the intervention and reductions in malaria that was uniform across age groups. Strategies have been developed for searching for effect modification while controlling the familywise Type I error rate for having conducted such a search (Hsu, Zubizarreta, Small and Rosenbaum, 2015; Lee, Small and Rosenbaum, 2017).

Sensitivity analysis can only quantify how large a bias in treatment assignment it would take to change the conclusions of an observational study compared to assuming randomized assignment – it can't tell us whether such bias in present. Chapters 7 and 8 of *Observation and Experiment* are about strategies for testing and understanding whether biases in treatment assignment are present. Chapter 7 presents the strategy of developing an "elaborate theory" and checking whether its consequences hold. Cochran, citing Fisher, said that one should "make your theories elaborate," meaning in Cochran's words, "when constructing a causal hypothesis, one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold." For example, Doll and Hill (1966) studied the effect of smoking on heart disease and asserted an elaborate theory – if smoking causes heart disease, not only should smokers have a higher death rate from heart disease than nonsmokers, but more specifically, light smokers should have mortality somewhere between that of nonsmokers and heavy smokers and people who quit smoking should have risks between those of nonsmokers and heavy smokers (it is not clear what to expect when comparing continuing light smokers to people who quit heavy smoking). If we were to find say that heavy smokers and light smokers have the same risk of heart disease even though they are both higher than nonsmokers, it might make us less confident that smoking really causes heart disease rather than that there is an unmeasured difference between smokers and nonsmokers; whereas if we had only compared smokers and nonsmokers, we might be convinced that smoking causes heart disease. In fact, Doll and Hill found that heavy smokers have higher risk of heart disease than light smokers and all of the other consequences of the elaborate theory hold. While finding that a consequence of the elaborate theory does not hold makes us less confident in the causal theory, finding that all consequences of the elaborate theory hold might not make us more confident in the causal theory if a bias could make all the consequences hold, e.g., there could be a genetic variant that causes heart disease and makes one more likely to smoke and also more likely to smoke heavily given that one smokes.

Rosenbaum discusses strategies for constructing elaborate theories that increase confidence when all the consequences hold. One is to specify consequences and studies for testing them that would be affected by different biases. Rosenbaum gives the example that an elaborate theory for smoking causing lung cancer might predict that "(i) smokers will develop lung cancer more often than nonsmokers in observational studies of people, (ii) laboratory animals experimentally exposed to tars in cigarettes will develop cancer, (iii) the autopsied lungs of smokers who died of something other than lung cancer will exhibit cellular damage similar to that of individuals who died of lung cancer and unlike the lungs of nonsmokers. These very different types of research each have weakness – smoking is not randomized and mice are not people – so each comparison may be unconvincing on its own, but agreement between studies with very different weaknesses may be compelling."

Different observational studies that are subject to different sources of bias can strengthen evidence when they agree. An observational study that removes a routine source of bias even while other sources of bias might be present can provide useful information, particularly in combination with other investigations. For example, consider the question of whether stress causes cardiac ischemia (restriction in blood supply to the heart which can lead to heart attacks and coronary death)? (The discussion of this example is drawn from Rosenbaum, 2001). Comparing individuals with high vs. low levels of stress contains important potential biases, for example highly stressed people may have poorer diet and get less exercise. Gullette et al. (1997) removed these routine sources of bias by comparing individuals to themselves – 132 patients with coronary disease were monitored for 48 hours using ambulatory electrocardiography devices to measure ischemia and a mood diary to measure stress. The frequency of tension, sadness or frustration (moods associated with stress) was compared in periods of ischemia to the preceding period in which there was not ischemia. Gullette et al. found more than a doubling of odds of tension, sadness and frustration in the periods of ischemia, suggesting that stress causes ischemia. Although this study removes bias from differences in diet and exercise between the high vs. low stress groups, it could have other sources of bias, e.g., subtle biological changes undetected by electrocardiography cause changes in both mood and ischemia. An observational study that like Gullette et al.'s study removed the routine sources of bias from comparing high stress vs. low stress individuals who might differ in long term diet and exercise but in a different way than Gullettte et al.'s study is Trichopoulos, Katsouyanni, Zavitsanos, Tzonou, and Dalla-Vorgia (1983), who compared coronary mortality in Athens in the days following the 1981 earthquake of magnitude 6.7 on the Richter scale with immediately adjacent time periods and the corresponding time periods in 1980 and 1982. Claiming that "the psychological stress [following the earthquake] was unquestionable, intense, and general," they found elevated rates of coronary mortality immediately following the earthquake when compared with the other time periods. This study could be biased by the earthquake changing the environment and individual behavior in ways besides stress that increase cardiac ischemia, e.g., responding to the damage caused by an earthquake may require physical exertion, and perhaps exertion and not stress causes ischemia. However, evidence for stress causing ischemia might be seen as strengthened by the combination of both Gullette et al.'s study and Trichopoulos et al.'s study finding that stress is associated with ischemia because each study is subject to its own bias but the biases are different from each other and both remove the routine source of bias from comparing high stress vs. low stress people. Mervyn Susser

said that evidence is strengthened when "diverse approaches produce similar results," and Rosenbaum says that this is particularly true "when these diverse approaches suffer from diverse weaknesses that are unlikely to align to produce similar results in the absence of a treatment effect." Diverse and independent approaches can be constructed within one study using Rosenbaum's recently developed tool of evidence factors that is discussed in Chapter 7; see Zhang, Small, Lorch, Srinivas and Rosenbaum (2011) for an application of evidence factors.

The way of elaborating a theory that most directly confronts potential biases is to make it so that we expect that some of the consequences of the elaborate theory will not hold if there is bias. As an example, Rosenbaum considers Wintemute, Wright, Drake and Beaumont's (2002) study of the effect on crime rates of a 1991 California law that restricts handgun purchases by people convicted of certain violent misdemeanors. Wintemute et al. compared crime rates of people convicted of violent misdemeanors who attempted to purchase a handgun in 1991 and were denied because of the new law vs. those who attempted to purchase a handgun in 1989 and 1990 who were able to make the purchase since the new law was not yet in effect (the outcome for each individual was the crime rate in the three years subsequent to the attempted purchase). A potential bias is that the denials all occurred in a later year than the approvals. Changes in the unemployment rate might affect whether a person is inclined to commit a crime. Rosenbaum says, "We want an elaboration of the causal theory so that a sour economy predicts one thing will happen but an effect of handgun restrictions predicts something else will happen." Such an elaboration is available. Restriction on handguns should reduce specifically crimes for which possession of a handgun is relevant. On the other hand, a rise in the unemployment rate in a particular year might increase the rate of crimes committed for monetary gain, but that tendency seems unlikely to be restricted to crimes for which a gun is relevant. Wintemute et al. found that the group denied purchase of a handgun in 1991 had lower rates of violent crime but not of nonviolent crime, a result Rosenbaum says is "easier to explain as an effect of restrictions on gun purchases, harder to explain in terms of a sour economy." Examining whether there is a difference in nonviolent crimes between the group affected by the law vs. the group not affected is a direct test for whether there is bias in treatment assignment by checking whether the treatment is associated with an outcome for which the treatment is known not to have an effect. In Wintemute et al.'s case, we expect the test to have power against a bias of concern, that a sour economy will increase crime rates in general. In other studies, we might be less certain that the test will have power against a bias of concern but it is still worth doing the test as if the test finds evidence of bias, we have learned that we should question the assumption that treatment assignment resembles randomized assignment. For example, Trichopoulos et al., in the aforementioned study of the effect of the Athens earthquake on coronary mortality, tested whether the earthquake was associated with cancer mortality. One bias that we were concerned about is that the damage caused by an earthquake requires physical exertion and that perhaps exertion and not stress causes coronary mortality. If increases in physical exertion do not increase cancer mortality, then the test does not have power to detect this alternative; nevertheless the test has power to detect other sources of bias and is worth doing. The Chinese proverb says, "He who asks a question is a fool for five minutes; he who does not ask a question remains a fool forever." The distinction between the test in Wintemute et al.'s study, which has power against a

specific bias of concern vs. the test in Trichopoulos et al.'s study which may not have power against a specific bias of concern but has power against other, not clearly specified biases is related to the distinction between active steps to detect hidden bias and tests of coherence made by Rosenbaum in his earlier book *Observational Studies* (Chapter 6-8 vs. Chapter 9).

Much of Rosenbaum's focus in *Observation and Experiment* is on how to address the concern in an observational study that there are unmeasured differences between the treatment and control groups. To make valid causal inferences, measured differences also need to be dealt with. Rosenbaum prefers matching to deal with measured differences comparing outcomes among treated and control units in matched sets in which the treated and control units have similar measured covariates, most simply comparing outcomes in matched pairs of treated and control units (Chapter 11). Matching has a number of attractive features as a way of adjusting for measured covariates:

1. Matching is easily understood and transparent. The comparison between the distribution of measured covariates in matched treated and control units can be presented in a table like a "Table 1" of a randomized trial or in Love plots. Subject matter experts can discuss, are the matched groups sufficiently similar that we would not expect a substantial difference between the groups in the absence of a treatment effect? Are there unmeasured covariates such that we would expect a substantial difference between the groups in the absence of a treatment effect even if the groups are well matched on the measured covariates? Rosenbaum says, "Ornate adjustments for observed covariates can, and often do, inhibit critical discussion...To be compelling, an observational study must speak to the issues that make observational evidence debatable; it must engage, not avoid, the debate." My experience has been that by looking along with my scientific collaborators at a "Table 1" of differences between the treatment and control groups before and after matching, we have often discovered differences we did not expect that provided insight into possible unmeasured confounders and understood that some variables were measured in different ways than we thought;

2. Matching facilitates blinding like in a randomized trial. Rosenbaum says, "A matched comparison should include one primary analysis [or a few primary analyses with appropriate adjustment for multiple inferences], selected during the design of the study before any outcomes were examined...An investigator who performs many complex analyses exercises enormous choice in the presentation of scientific findings, and a simple, prespecified primary analysis is intended to place a sharp limit on this sort of spin-doctoring." In a recent observational study of the effect of playing high school football on later life depression and cognitive functioning (Deshpande et al., 2017), we posted a protocol to ArXiv with the matching "Table 1" that compares matched treated and control units on measured covariates and our prespecified primary analysis before examining the outcome data. The investigator can work hard at balancing the covariates without looking at the outcome data, perhaps rejecting an initial match as inadequate because a covariate is not balanced; once adequate balance has been achieved, the analysis can be conducted in a simple, nonparametric way such as using Wilcoxon's signed rank test with matched pairs. In contrast when using regression to adjust for measured covariates, one might work harder at diagnosing model misfit

when an initial model suggests a treatment effect opposite to the expected direction than when it is in the expected direction.

3. Matching facilitates understanding overlap are there some treated (control) units for which we don't have sufficient comparable control (treated) units to make a comparison and estimate the treatment effect? If for a treated unit we can't find a control unit that is somewhat similar on measured covariates, we can't hope to estimate the treatment effect for that unit without extrapolation. Fogarty, Mikkelsen, Gaieski and Small (2016) provide a method for defining an interpretable subpopulation for which a matched comparison can be conducted without extrapolating with respect to important variables

4. Matching provides a framework within which qualitative and quantitative research can usefully interact within a single investigation (Rosenbaum and Silber, 2001). A few matched pairs can be closely examined and narrative accounts made that might reveal why one unit in the pair chose treatment and the other control in spite of having similar measured covariates, i.e., reveal an unmeasured confounder.

   Rosenbaum's book is accessible to readers of all backgrounds, but at the same time, contains much material of interest to experienced causal inference researchers that is different from his previous two books *Observational Studies* and *Design of Observational Studies*. An example is Chapter 12 which describes a new approach to controlling for biases from general dispositions (Rosenbaum, 2006). Consider a study of the effect of wearing a helmet on injury severity from a bike fall and comparing Harry who didn't wear a helmet and Sally who did wear a helmet. Rosenbaum says, "We often attribute a particular choice made by a person to a disposition of that person to make choices in a particular way. Harry does not wear seatbelts and texts while driving, often tailgates, and does not wear his helmet when cycling because Harry is a reckless person. Sally wears seatbelts and never texts when driving, never tailgates and wears a helmet when cycling because she is a cautious person...Your data record that Harry and Sally each fell from their bicycles and Harry's injuries were more severe, but the data do not record that Sally hit a bump at a slow speed and landed in the grass, while Harry sped through a red signal, dodged a car, lost control and slammed into a lamppost." The disposition of recklessness is a confounder that we need to control for. But matching for observed reckless habits may not be quite enough to compensate for a general disposition of recklessness – there are many manifestations of recklessness and our data may only record a few. Rosenbaum proposes a new way to put these few recorded manifestations to use. Consider David who, "never texts while driving that would be totally irresponsible but he skips the helmet while cycling because he likes the wind in his hair" and Debbie who, "never skips the helmet while cycling that would be totally irresponsible but she texts while driving because she is good at multitasking." The usual approach would be to pair David to a helmeted control who never texts while driving like Sally; then David resembles his control in terms of texting. But Sally appears to be a less reckless person than David and this may manifest itself in many unmeasured ways like cycling speed and stopping at red traffic signals. Instead of pairing David to Sally, Rosenbaum proposes pairing David to Debbie despite their visible difference on covariates (David never texts while driving but

Debbie does) because their overall visible behavior (David never texts while driving but wears a helmet when cycling and Debbie does the opposite) indicates a middling disposition toward recklessness and they may be similar in terms of cycling speed and stopping at red traffic signals. Rosenbaum shows that under a Rasch model for behaviors, comparing people like David to people like Debbie produces an unbiased estimate of the causal effect of wearing a helmet. There seems to be room for interesting future research here. How can the evidence from this type of comparison best be integrated with traditional comparisons of people who are the same on measured covariates? The Rasch model is one type of item response theory model for measuring latent traits like recklessness could other item response theory models be useful in designing observational studies?

The genius of the randomized experiment is that it provides a mechanical way to infer the causal effect of a treatment as long as a certain procedure (randomization of treatment assignment) is followed. Rosenbaum stresses that there is no mechanical way to infer the causal effect of a treatment in an observational study because there is no mechanical way to remove the possibility of unmeasured confounding without assumptions. Instead, he presents considerations and strategies for examining whether unmeasured confounding is present and designing observational studies that have reduced sensitivity to unmeasured confounding. Good use of these considerations and strategies requires thinking and knowledge (or collaboration with others knowledgeable) about the subject matter.

In summary, *Observation and Experiment* is a treasure trove of considerations and strategies for making causal inferences from observational studies and experiments. The book is a joy to read and contains interesting material for readers at all levels of experience with causal inference.

References

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173-203.

Deshpande, S.K., Hasegawa, R.B., Rabinowitz, A.R., Whyte, J., Roan, C.L., Tabatabaei, A., Baiocchi, M., Karlawish, J.H., Master, C.L. and Small, D.S. (2017). Association of playing high school football with cognition and mental health later in life. *JAMA Neurology*, 74, 909-918.

Doll, R. and Hill, A.B. (1966). Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. In: *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases*, W. Haenszel, ed., U.S. National Cancer Institute Monograph 19, U.S. Depatment of Health, Education and Welfare: Washington, D.C., pp. 205-268.

Dorn, H. F. (1959). Tobacco consumption and mortality from cancer and other diseases. *Public health reports*, 74, 581-593.

Fisher, R. A. (1958). Lung cancer and cigarettes?. *Nature*, 182, 108.

Fogarty, C.B., Mikkelsen, M.E., Gaieski, D.F. and Small, D.S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111, 447-458.

Frost, R. (1916). The Road Not Taken. In *Mountain Interval* by R. Frost. Henry Holt and Company. New York.

Gullette, E., Blumenthal, J., Babyak, M., Jiang, W., Waugh, R., Frid, D., O'Connor, C., Morris, J. and Krantz, D. (1997). Effects of mental stress on myocardial ischemia during daily life. *Journal of the American Medical Association*, 277, 15211526.

Hsu, J.Y., Small, D.S. and Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108, 135-148.

Hsu, J.Y., Zubizarreta, J.R., Small, D.S. and Rosenbaum, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102, 767-782.

Imbens, G.W., Rubin, D.B. and Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91, 778-794.

Lee, K., Small, D.S. and Rosenbaum, P.R. (2017). A new, powerful approach to the study of effect modification in observational studies. arXiv preprint:1702.00525.

Oreopoulos, P. (2003). The long-run consequences of living in a poor neighborhood. *Quarterly Journal of Economics*, 118, 1533-1575.

Rosenbaum, J.E. (1995). Changing the geography of opportunity by expanding residential choice: Lessons from the Gautreaux program. *Housing Policy Debate*, 6, 231-269.

Rosenbaum, P. R. (2001). Replicating effects and biases. *The American Statistician*, 55, 223-227.

Rosenbaum, P.R. (2002). *Observational studies*. Springer: New York.

Rosenbaum, P. R. (2006). Differential effects and generic biases in observational studies. *Biometrika*, 93, 573-586.

Rosenbaum, P.R. and Silber, J.H. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics*, 2, 217-232.

Trichopoulos, D., Katsouyanni, K., Zavitsanos, X., Tzonou, A. and Dalla-Vorgia, P. (1983). Psychological stress and fatal heart attack: The Athens 1981 earthquake natural experiment. *Lancet*, 321, 441444.

Votruba, M.E. and Kling, J.R. (2009). Effects of neighborhood characteristics on the mortality of black male youth: Evidence from Gautreaux, Chicago. *Social Science & Medicine*, 68, 814-823.

Wintemute, G.J., Wright, M.A., Drake, C.M. and Beaumont, J.J. (2001). Subsequent criminal activity among violent misdemeanants who seek to purchase handguns: risk

factors and effectiveness of denying handgun purchase. *Journal of the American Medical Association*, 285, 1019-1026.

Zhang, K., Small, D. S., Lorch, S., Srinivas, S., and Rosenbaum, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association*, 106, 511-524.

Zubizarreta, J.R., Small, D.S. and Rosenbaum, P.R. (2014). Isolation in the construction of natural experiments. *Annals of Applied Statistics*, 8, 2096-2121.