11

Observational Causal Studies

> Shallow men believe in luck. Strong men believe in cause and effect
> – *Ralph Waldo Emerson*

Why you care: *Randomized controlled experiments are the gold standard for establishing causality, but sometimes running such an experiment is not possible. Given that organizations are collecting massive amounts of data, there are observational causal studies that can be used to assess causality, although with lower levels of trust. Understanding the space of possible designs and common pitfalls can be useful if an online controlled experiment is not possible.*

When Controlled Experiments Are Not Possible

What is the impact on product engagement if a user switches their phone from an iPhone to a Samsung? How many users come back if we forcibly sign them out? What happens to revenue if coupon codes are introduced as part of the business model? For all these questions, the goal is to measure the causal impact for a change, which requires comparing the outcome of a treated population to the outcome for an untreated population. The "basic identity of causal inference" (Varian 2016) is:

$$\begin{aligned}&\text{Outcome for treated} - \text{Outcome for untreated}\\ &= [\text{Outcome for treated} - \text{Outcome for treated if not treated}]\\ &\quad + [\text{Outcome for treated if not treated} - \text{Outcome for untreated}]\\ &= \text{Impact of Treatment on treated} + \text{Selection bias}\end{aligned}$$

and shows that the comparison of the actual impact (what happens to the treated population) compared to the counterfactual (what would have

happened if they had not been treated) is the critical concept for establishing causality (Angrist and Pischke 2009, Neyman 1923, Rubin 1974, Varian 2016, Shadish, Cook and Campbell 2001).

Controlled experiments are the gold standard for assessing causality because, with random assignment of units to variants, the first term is the observed difference between Treatment and Control and the second term has an expected value of zero.

However, sometimes you cannot run a properly controlled experiment. These situations include:

- When the causal action to be tested is not under the control of the organization. For example, you may want to understand how a user's behavior changes when they change their phone from an iPhone to a Samsung Galaxy phone. Even if you are Samsung with some levers to incent users to switch that can be randomized, generally you are not in control of users' choices here and paying people to switch biases the results.
- When there are too few units. For example, in a Merger and Acquisition (M&A) scenario, there is a single event that happens (or not) and estimating the counter-factual is extremely hard.
- When establishing a Control may incur too large an opportunity cost since they do not receive the Treatment (Varian 2016). For example, randomized experiments can be costly for rare events, such as establishing the impact of running ads during the Superbowl (Stephens-Davidowitz, Varian and Smith 2017), or when the desired OEC takes too long to measure, such as returning to a website to purchase a new car five years after the current car purchase.
- When the change is expensive relative to the perceived value. Some experiments are run to try to better understand relationships. For example, how many users will churn if you forcibly sign out all users after some time period? Or, what if you don't display ads on a search engine such as Bing or Google?
- When the desired unit of randomization cannot be properly randomized. When assessing the value of TV ads, it is practically impossible to randomize by viewers. The alternative of using Designated Market Areas (DMAs) (Wikipedia contributors, Multiple Comparisons problem 2019), results in far fewer units (e.g., about 210 in the US) and hence low statistical power, even when using techniques such as pairing.
- When what is being tested is unethical or illegal, such as withholding medical treatments that are believed to be beneficial.

In the above situations, often the best approach is to estimate the effects using multiple methods that are lower in the hierarchy of evidence, that is, answering the question using multiple methods, including small-scale user experience

studies, surveys, and observational studies. See Chapter 10 for an introduction to several other techniques.

Our focus in this chapter is on estimating the causal effect from observational studies, which we will call *observational causal studies*. Some books, such as Shadish et al. (2001), use the term *observational (causal) studies* to refer to studies where there is no unit manipulation, and the term *quasi-experimental designs* to studies where units are assigned to variants, but the assignment is not random. For additional information, please see Varian (2016) and Angrist and Pischke (2009, 2014). Note that we differentiate an observational causal study from the more general observational, or retrospective, data analyses. While both are run on historical log data, the goal in an observational causal study is to try to get as close to a causal result as possible, while retrospective data analyses, as discussed in Chapter 10, have different goals, ranging from summarizing distributions, seeing how common certain behavioral patterns are, analyzing possible metrics, and looking for interesting patterns that may suggest hypotheses to be tested in controlled experiments.

Designs for Observational Causal Studies

In observational causal studies, the challenges are:

- How to construct Control and Treatment groups for comparison.
- How to model the impact given those Control and Treatment groups.

Interrupted Time Series

Interrupted Time Series (ITS) is a quasi-experimental design, where you can control the change within your system, but you cannot randomize the Treatment to have a proper Control and Treatment. Instead, you use the same population for Control and Treatment, and you vary what the population experiences over time.

Specifically, it uses multiple measurements over time, before an intervention, to create a model that can provide an estimate for the metric of interest after the intervention – a counterfactual. After the intervention, multiple measurements are taken, and the Treatment effect is estimated as the average difference between the actual values for the metric of interest and those predicted by the model (Charles and Melvin 2004, 130). One extension to the simple ITS is to introduce a Treatment and then reverse it, optionally repeating this procedure multiple times. For example, the effect of police

helicopter surveillance on home *burglaries* was estimated using multiple Treatment interventions because over several months, surveillance was implemented and withdrawn several times. Each time that helicopter surveillance was implemented, the number of burglaries decreased; each time surveillance was removed, the number of burglaries increased (Charles and Melvin 2004). In an online setting, a similar example is to understand the impact of online advertising on search-related site visits. Note that sophisticated modeling may be necessary to infer the impact, with an online example of ITS being Bayesian Structural Time Series analysis (Charles and Melvin 2004).

One common issue with observational causal studies is ensuring that you are not attributing an effect to a change when in fact there is some confounding effect. The most common confounds for ITS are time-based effects as the comparisons are made across different points of time. Seasonality is the obvious example, but other underlying system changes can also confound. Changing back and forth multiple times will help reduce the likelihood of that. The other concern when using ITS is on the user experience: Will the user notice their experience flipping back and forth? If so, then that lack of consistency may irritate or frustrate the user in a way that the effect may not be due to the change but rather the inconsistency.

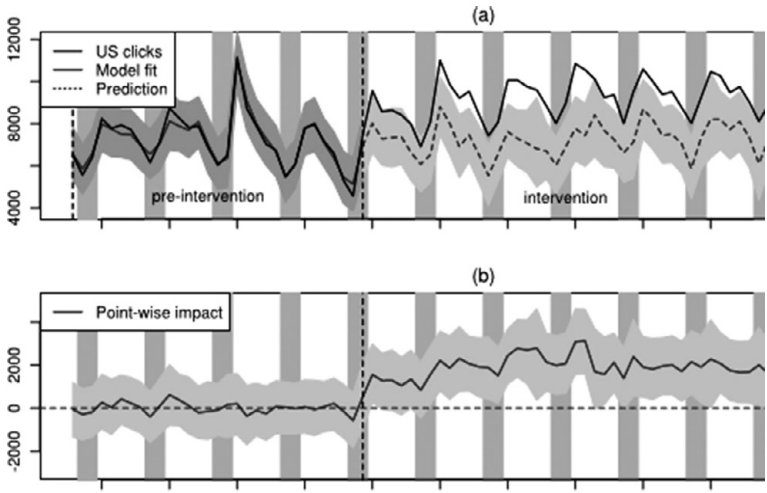

Figure 11.1 Interrupted Time Series using Bayesian Structural Time Series (Charles and Melvin 2004). (a) shows the model fit in the pre-intervention period and the actual observed metric in the solid line, with the dashed line the predicted counterfactual. The x-axis is days with shaded vertical bars indicating weekends. (b) shows the delta between the actual and the prediction; if the model is good, then it is an estimate of the Treatment effect. Weekends are shaded in grey

Interleaved Experiments

Interleaved experiment design is a common design used to evaluate ranking algorithm changes, such as in search engines or search at a website (Chapelle et al. 2012, Radlinski and Craswell 2013). In an interleaved experiment, you have two ranking algorithms, X and Y. Algorithm X would show results x_1, x_2, $\ldots x_n$ in that order, and algorithm Y would show y_1, y_2, $\ldots y_n$. An interleaved experiment would intersperse results mixed together, e.g. x_1, y_1, x_2, y_2, $\ldots x_n$, y_n with duplicate results removed. One way to evaluate the algorithms would be to compare the click-through rate on results from the two algorithms. While this design is a powerful experiment design, it is limited in its applicability because the results must be homogenous. If, as is common, the first result takes up more space, or impacts the other areas of the page, then complexities arise.

Regression Discontinuity Design

Regression Discontinuity Design (RDD) is a methodology that can be used whenever there is a clear threshold that identifies the Treatment population. Based on that threshold, we can reduce selection bias by identifying the population that is just below the threshold as Control and compared to the population that is just above the threshold as Treatment.

For example, when a scholarship is given, the near-winners are easily identified (Thistlewaite and Campbell 1960). If a scholarship is given for an 80% grade, then the Treatment group that received grades just above 80% is assumed to be similar to the Control group that received grades just below 80%. The assumption is violated when participants can impact their Treatment; for example, if the Treatment is applied to a passing grade, but students are able to convince their teachers to "mercy pass" them (McCrary 2008). An example using RDD is in assessing the impact of drinking on deaths: Americans over 21 can drink legally, so we can look at deaths by birthday, shown in Figure 11.2. The "Mortality risk shoots up on and immediately following a twenty-first birthday … about 100 deaths to a baseline level of about 150 per day. The age-21 spike doesn't seem to be a generic party-hardy birthday effect. If this spike reflects birthday partying alone, we should expect to see deaths shoot up after the twentieth and twenty-second birthdays as well, but that doesn't happen" (Angrist and Pischke 2014).

As in the above example, one key issue is again confounding factors. In RDD, the threshold discontinuity may be contaminated by other factors that share the same threshold. For example, a study of the impact of alcohol that chooses the legal age of 21 as the threshold may be contaminated by the fact that this is also the threshold for legal gambling.

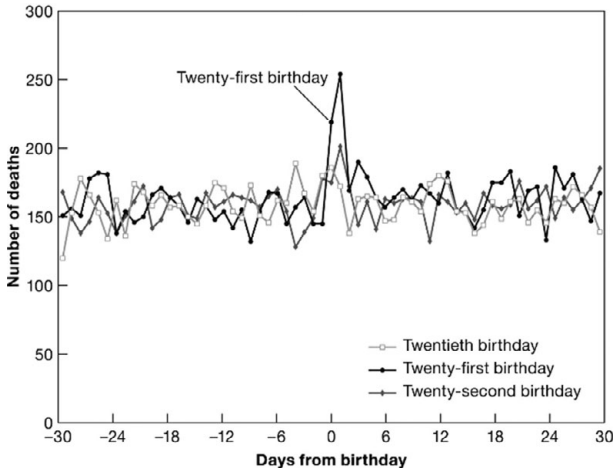

Figure 11.2 Deaths vs. Days from birthday for 20th, 21st, and 22nd birthday (Angrist and Pischke 2014)

RDD most commonly applies when there is an algorithm that generates a score, and something happens based on a threshold of that score. Note that when this happens in software, while one option is to use RDD, this is also a scenario that also easily lends itself to randomized controlled experiments, or some hybrid of the two (Owen and Varian 2018).

Instrumented Variables (IV) and Natural Experiments

Instrumental Variables (IV) is a technique that tries to approximate random assignment. Specifically, the goal is to identify an Instrument that allows us to approximate random assignment (this happens organically in a natural experiment) (Angrist and Pischke 2014, Pearl 2009).

For example, to analyze difference in earnings between veterans and non-veterans, the Vietnam war draft lottery resembles random assignment of individuals into the military; charter school seats are allocated by lottery and can thus be a good IV for some studies. In both examples, the lottery does not guarantee attendance but has a large impact on attendance. A two-stage least-squares regression model is then commonly used to estimate the effect.

Sometimes, natural experiments that are "as good as random" can occur. In medicine, monozygotic twins allow running twin studies as natural experiments (Harden et al. 2008, McGue 2014). Online, when studying social or peer networks, running controlled experiments on members can be challenging as

the effect may not be constrained to the Treatment population due to member-to-member communications. However, notification queues and the message delivery order are types of natural experiments that can be leveraged to understand the impact of notifications on engagement, for example Tutterow and Saint-Jacques (2019).

Propensity Score Matching

Another class of approaches here is to construct comparable Control and Treatment populations, often by segmenting the users by common confounds, in something akin to stratified sampling. The idea is to ensure that the comparison between Control and Treatment population is not due to population mix changes. For example, if we are examining an exogenous change of the impact of users changing from Windows to iOS, we want to ensure that we are not measuring a demographic difference in the population.

We can take this approach further by moving to propensity score matching (PSM) that, instead of matching units on covariates, matches on a single number: a constructed propensity score (Rosenbaum and Rubin 1983, Imbens and Rubin 2015). This approach has been used in the online space, for example for evaluating the impact of online ad campaigns (Chan et al. 2010). The key concern about PSM is that only observed covariates are accounted for; unaccounted factors may result in hidden biases. Judea Pearl (2009, 352) wrote "Rosenbaum and Rubin ... were very clear in warning practitioners that propensity scores work only under 'strong ignorability' conditions. However, what they failed to realize is that it is not enough to warn people against dangers they cannot recognize." King and Nielsen (2018) claim that PSM "often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias."

For all of these methods, the key concern is confounding factors.

Difference in Differences

Many of the methods above focus on how to identify a Control group that is as similar to a Treatment group as possible. Given that identification, one method to measure the effect of the Treatment is difference in differences (DD or DID) that, assuming common trends, assigns the difference in difference to the Treatment. In particular, the groups "may differ in the absence of Treatment yet move in parallel" (Angrist and Pischke 2014).

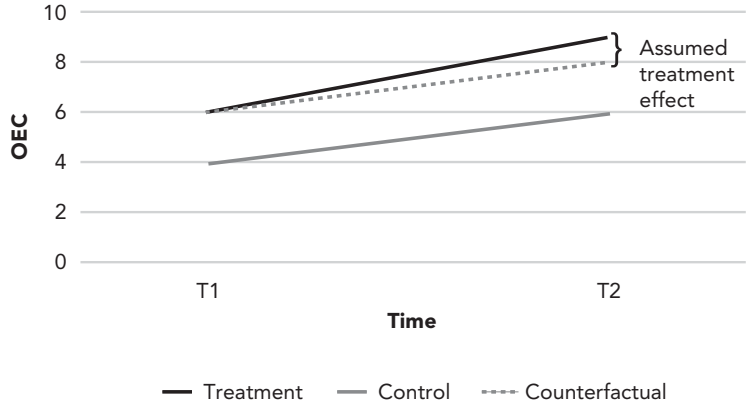

Figure 11.3 Difference in differences

Geographically based experiments commonly use this technique. You want to understand the impact of TV advertising on driving user acquisition, engagement, and retention. You run TV ads in one DMA and compare it to another DMA. For example, as shown in Figure 11.3, a change is made at time T_1 to the Treatment group. Measurements are taken for both the Treatment and Control just before T_1, and at a later point T_2. The difference in the metrics of interest, such as the OEC, between the two periods in the Control group are assumed to capture the external factors (e.g., seasonality, economic strength, inflation) and thus present the counterfactual of what would have happened to the Treatment group. The Treatment effect is estimated as the difference in a metric of interest minus the difference in Control for that metric over the same period.

Note that this method can also be applied even when you do not make the change, and the change happens exogenously. For example, when a change was made to the minimum wage in New Jersey, researchers who wanted to study its impact on employment levels in fast-food restaurants, compared it to eastern Pennsylvania, which matched on many characteristics (Card and Krueger 1994).

Pitfalls

Although observational causal studies are sometimes your best option, they have many pitfalls that you should be aware of (see also Newcomer et al. (2015) for a more exhaustive list). As mentioned above, the main pitfall,

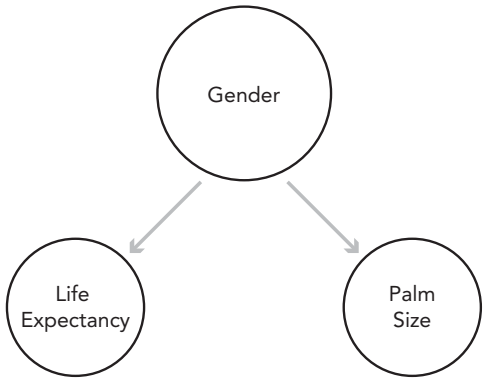

Figure 11.4 Instead of Palm Size predicting Life Expectancy, Gender is the common cause that predicts both

regardless of method, in conducting observational causal studies is unanticipated confounds that can impact both the measured effect, as well as the attribution of causality to the change of interest. Because of these confounds, observational causal studies require a great deal of care to yield trustworthy results, with many examples of refuted observational causal studies (see *Sidebar: Refuted Observational Causal Studies* later in this chapter and Chapter 17 for a few).

One common type of confound is an unrecognized **common cause**. For example, in humans, palm size has a strong correlation with life expectancy: on average the smaller your palm, the longer you will live. However, the common cause of smaller palms and longer life expectancy is gender: women have smaller palms and live longer on average (about six years in the US).

As another example, for many products, including Microsoft Office 365, users that see more errors typically churn *less*! But do not try to show more errors expecting to reduce churn, as this correlation is due to a common cause: usage. Your heaviest users see more errors *and* churn at lower rates. It is not uncommon for feature owners to discover that users of their new feature churn at a lower rate, which implies that it is their feature that is reducing churn. Is it really the feature or (more likely) simply that heavy users churn less and are more likely to use more features? In these cases, to evaluate whether the new feature indeed reduces churn, run a controlled experiment (and analyze new and heavy users separately).

Another pitfall to be aware of are **spurious or deceptive correlations**. Deceptive correlations can be caused by strong outliers, for example as in

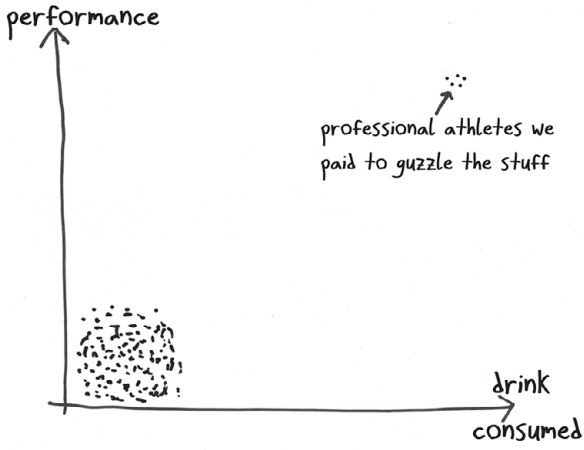

Figure 11.5 Deceptive correlation of athletic performance and amount of energy drink consumed. Correlation does not imply causation!

Figure 11.5, where a marketing company can claim that their energy drink is highly correlated with athletic performance and imply a causal relationship: drink our energy product and your athletic performance will improve (Orlin 2016).

Spurious correlations can almost always be found (Vigen 2018). When we test many hypotheses and when we do not have the intuition to reject a causal claim as we have in the above example, we may believe it. For example, if someone told you that they found a factor that had a strong correlation (r=0.86) with people killed by venomous spiders, you might be tempted to act on this information. Yet when you realize that the deaths are correlated with word length in the National Spelling Bee test, as shown in Figure 11.6, you quickly reject the request to shorten the word length in the National Spelling Bee as irrational.

Even when care is taken, there is never a guarantee that there is not some other factor not included in the observational causal study that may impact the results. Quasi-experimental methods, which attempt to derive a counterfactual to compare to and therefore establish causality, simply require making many assumptions, any of which can be wrong, and some assumptions are implicit. Incorrect assumptions can lead to a lack of internal validity but depending on the assumptions and how limiting they are, they can also impact the external validity of the study. While building intuition, as discussed in Chapter 1, can help improve the quality of assumptions, intuition will not mitigate all possible problems. Thus, the scientific gold standard for establishing causality is still the controlled experiment.

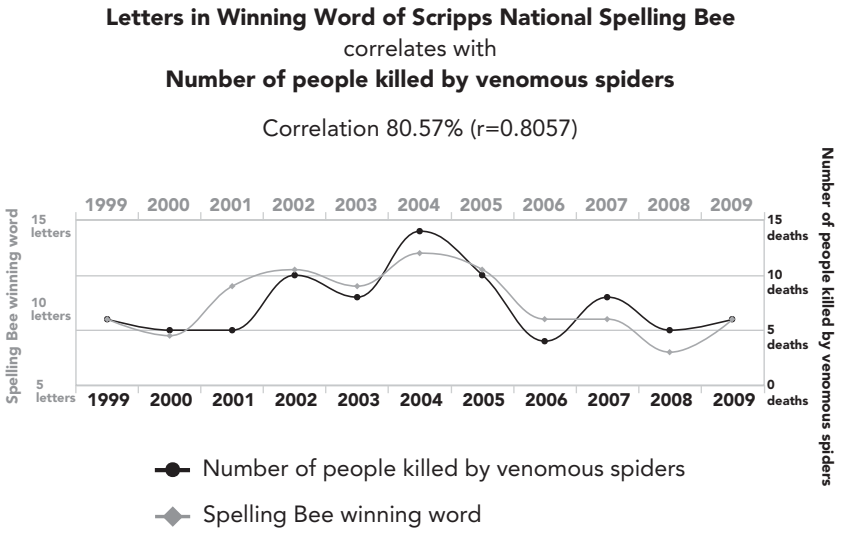

Figure 11.6 Spurious correlation of people killed by venomous spiders and word length in Scripps National Spelling Bee

SIDEBAR: Refuted Observational Causal Studies

Claiming causality from observational data (uncontrolled) requires multiple assumptions that are impossible to test and are easily violated. While many observational causal studies are later confirmed by randomized controlled experiments (Concato, Shah and Horwitz 2000), others are refuted. Ioannidis (2005) evaluated claims coming from highly cited studies; of six observational causal studies included in his study, five failed to replicate. Stanley Young and Alan Karr (2019) compared published results from medical hypotheses shown to be significant using observational causal studies (i.e., uncontrolled) with randomized clinical trials considered more reliable. Of 52 claims in 12 papers, none replicated in the randomized controlled trials. And in 5 of the 52 cases, the direction was statistically significant in the opposite direction of the observational causal study. Their conclusion: "Any claim coming from an observational study is most likely to be wrong."

One example from the online space is on how to measure the effectiveness of online advertising, in other words, whether online ads led to either increased brand activity or even user engagement. Observational causal studies are often required to measure the effect, since the intervention (the ad) and the effect (user sign-up or engagement) are typically on different sites and therefore different spheres of control. Lewis, Rao and Reiley (2011) compared the

effectiveness of online advertising as estimated by observational causal studies relative to the "gold standard" controlled experiments, finding that observational causal studies vastly overestimated the effect. Specifically, they ran three experiments.

First, advertisements (display ads) were shown to users, and the question was: What is the increase (lift) in the number of users who search using keywords related to the brand shown in the ad. Using several observational causal studies of 50 million users, including three regression analyses with Control variables, the estimated lift ranged from 871% to 1198%. This estimated lift is orders of magnitude higher from the lift of 5.4% measured via the controlled experiment. The confound is common cause of users visiting Yahoo! in the first place: Users who actively visit Yahoo! on a given day are much more likely to see the display ad and to perform a Yahoo! search. The ad exposure and the search behavior are highly positively correlated, but the display ads have very little causal impact on the searches.

Next, videos were shown to users, and the question was whether these would lead to increased activity. Users were recruited through Amazon Mechanical Turk, with half exposed to a 30-second video advertisement promoting Yahoo.com services (the Treatment), and half to a political video advertisement (the Control), and the goal was to measure whether there was increased activity on Yahoo! Two analyses were done: an observational causal study of the Treatment group before and after the exposure to the 30-second Yahoo! ads, and an experimental analysis comparing the activity of the two groups after seeing the ad. The observational causal study overstated the effects of the ad by 350%. Here, the common confound is that being active on Amazon Mechanical Turk on a given day increased the chance of participating in the experiment and being active on Yahoo!

Finally, an ad campaign was shown to users on Yahoo! with the goal of measuring whether users who saw the ad were more likely to sign up at the competitor's website on the day they saw the ad. The observational causal study compared users exposed to the ad on the day they saw the ad relative to the week before, while the experiment compared users who did not see the ad but visited Yahoo! on that day to the users who came to Yahoo! on the same day and saw the competitor ad. From the observational causal study, exposed users were more likely to sign up at the competitor's website the day they saw the ad compared to the week before. However, from the experiment, they observed a nearly identical lift. This result is similar to our previous discussion of churn and errors: More active users are simply more likely to do a broad range of activities. Using activity as a factor is typically important.

This is but one story and one comparison. A more recent comparison study also found that observational causal studies were less accurate than online controlled experiments (Gordon et al. 2018). We provide many more stories online at https://bit.ly/experimentGuideRefutedObservationalStudies, showing examples of unidentified common causes, time-sensitive confounds, population differences lack of external validity, and more. Should you need to do an observational causal study, please take care.