A whirlwind introduction to causal assumptions

JA Labrecque
Department of Epidemiology, Erasmus MC

Abstract

What is a causal effect?

I want to know whether aspirin cures headaches. To do this, I wait until the next time I have a headache. Then, I find another person with a headache who is as similar as possible to me. I take an aspirin and ask them not to take an aspirin. After an hour my headache goes away while the other person still has a headache. Can I say the aspirin cured my headache? No. Regardless of how similar this other person might be, what happened with their headache is not necessarily what would have happened with my headache had I not taken aspirin. To really know if the aspirin cured my headache, we would have to know what would have happened to me had I not taken the aspirin.

If you objected to me comparing myself to a different person, why then, in most epidemiologic studies, do we compare two *different* groups of people, one treated and one untreated? Are we not making the same mistake as in my aspirin example? You might argue that with two different groups you have a larger sample size or, possibly, the ability to adjust away differences between the groups. But neither of these considerations addresses the main problem from the previous paragraph: how do we know that what happened to the treated group is what would have happened to the untreated group had they been treated (and vice versa)? You cannot. We are in precisely the same situation as in the previous paragraph. We want to know how the outcome of a group of people would differ if we could observe them under two different conditions, treated and untreated, but that is impossible. All we can do is compare two different groups of people.

What I hope to have convinced you of is that comparing the outcome in a group of treated people to the outcome in different group of untreated people is *not the answer to the question we are interested in* when we are interested in causal effects. A causal question contrasts the outcome of the same group of people in two different states. Why then do we compare two different groups? It is the best we can do given that we can never observe the same people in two different states.

This is where causal inference comes in. We can measure the proportion of people in the treated group who get the outcome and the proportion of people in the untreated group who get the outcome and use these two values to get a risk difference:

$$\Pr(Y|T=1)-\Pr(Y|T=0)$$

$\Pr(Y|T=1)$ is the probability of Y among people with T=1, people who were, in fact, treated. $\Pr(Y|T=0)$ is the probability of Y among people with T=0, people who were, in fact, untreated. The vertical bar '|' can be thought of as "among people with". When you see '|' you know we are looking in a subgroup of people.

What we want to know, instead, is:

$$\Pr(Y^{t=1})-\Pr(Y^{t=0})$$

We use superscript to indicate that we are setting someone to receive a specific treatment even if, in fact, that is not what they received. Pr($Y^{t=1}$) is how we would write out the proportion of people who would have had outcome Y had we made them all take treatment *t=1*. Notice there are no vertical bars in this expression. We are really asking about the difference in proportion of people would get the outcome in everyone we are interested in was set to be treated, t=1, and the proportion of people who would get the outcome if everyone was set to be untreated, t=0.

The simplest way to think about what causal inference does is that it clarifies or clearly states the conditions under which the quantity we can observe is equal to the one we want to know. It tells us when we can believe the following:

$$\Pr(Y|T = 1) - \Pr(Y|T = 0) = \Pr(Y^{t=1}) - \Pr(Y^{t=0})$$

In the next section, we will go step by step to see how we can go from the left side of the equation to the right side.

The setup: what is the causal effect of Program T on infant mortality?

Let us imagine we want to know the causal effect of a program we will call Program T on infant mortality. Half of the hospitals in our hypothetical country use Program T while the other half do not. We have measured the average infant mortality both in hospitals that used Program T, 6.9/1,000, and those that did not, 7.6/1,000. These values can be found in the white half of Table 1 which represents the observable world. The grey half of the table are counterfactual quantities which we will need to fill in cell by cell using a combination of the observed data and causal assumptions. Once we have filled in the dark gray cells, we can take the average of each column to get the values at the bottom of table which we can contrast to get a causal effect. A quick word of warning that language can get a little tricky when talking about, for example, what outcome we would have observed among hospitals that used Program T had we made them not use Program. Parsing these sentences slowly may help.

Table 1—An outline of the basic setup of our example. The white half of the table is data from the observed world and the grey half is data from the counterfactual world. Each cell has the require value written in words and in statistical notation. The two values in the bottom row are those required to estimate a causal effect.

Observed world		Counterfactual world	
Exposure (T)	Outcome (Y)	$Y^{t=0}$	$Y^{t=1}$
0	$\Pr(Y\|T = 0) = 7.6$ The observed infant mortality in hospitals that did not use Program T	$\Pr(Y^{t=0}\|T = 0)$: The infant mortality we would have observed among hospitals that did not use Program T had they not used Program T	$\Pr(Y^{t=1}\|T = 0)$: The infant mortality we would have observed among hospitals that did not use Program T had they used Program T
1	$\Pr(Y\|T = 1) = 6.9$ The observed infant mortality in hospitals that did use Program T	$\Pr(Y^{t=0}\|T = 1)$: The infant mortality we would have observed among hospitals that used Program T had they not used Program T	$\Pr(Y^{t=1}\|T = 1)$: The infant mortality we would have observed among hospitals that did use Program T had they used Program T
Total (average of the two cells above)		$\Pr(Y^{t=0})$: The infant mortality we would have observed had all hospitals not used Program T	$\Pr(Y^{t=1})$: The infant mortality we would have observed had all hospitals used Program T

Causal assumptions when adjusting for confounding

Consistency

Read the text in Table 1 carefully to understand what each cell means. We have to use the information in the white part of the table to fill in the dark grey cells on the right side. Which two dark gray cells seem like a clear choice to fill in first? The top-left and bottom-right dark gray cells are both counterfactuals that it seems we have observed. The bottom right, for example, is the infant mortality we would have observed if we had set the hospitals that did use Program T to use Program T. These hospitals did, in actual fact, use Program T, so we did not even have to set them to that value. We can take the value we observed in hospitals that used Program T and plug it in.

It may look like no assumption is even being made here. We are plugging in a value for a counterfactual that we have observed directly. Or have we? What if, for example, the hospitals that used Program T did not all implement the program perfectly. Some hospitals applied the program later in the pregnancy than they were supposed to. Some hospitals may have omitted parts of the Program because they were too costly. If this were true, the observed infant mortality of 6.9/1000 likely does not correspond to what we would have observed had we made all hospitals follow Program T to the letter.

You might be inclined to say that we can estimate, instead, the effect of an imperfectly implemented version of Program T. As we have mentioned, however, different hospitals implemented Program T in different ways. To which imperfectly implemented version of Program T would our causal effect apply to?

The assumption we are discussing here is called the consistency assumption and is written more generally as $\Pr(\mathrm{Y}|\mathrm{T}=\mathrm{t})=\Pr(Y^{t}|T=t)$. In words, this is the observed risk among units (e.g., participants or hospitals) that were observed to get treatment *t* is equal to the risk we would have observed had we set them to get *T=t*. In even simpler words, when we say we want to ask a causal question about setting someone to treatment *t*, there is no ambiguity in what that means. For this reason, this is sometimes referred to as the "well-defined intervention" assumption. See the following references for more information on this. Let us proceed assuming this assumption is satisfied and therefore, we will plug in the corresponding values into Table 2.

Table 2—The consistency assumption is used to fill in the top-left and bottom-right dark grey cells.

Observed world		Counterfactual world	
Exposure (T)	Outcome (Y)	$Y^{t=0}$	$Y^{t=1}$
0	$\Pr(Y|T=0)=7.6$	$\Pr(Y^{t=0}|T=0)=7.6$	$\Pr(Y^{t=1}|T=0)$
1	$\Pr(Y|T=1)=6.9$	$\Pr(Y^{t=0}|T=1)$	$\Pr(Y^{t=1}|T=1)=6.9$
Total		$\Pr(Y^{t=0})$	$\Pr(Y^{t=1})$

Exchangeability

Filling in the last two cells is bit more challenging. What value could we possibly fill in, for example, for the bottom left dark-gray cell which is the infant mortality we would have observed among hospitals that used Program T had we set them to not use Program T? After all, this value is unobservable by definition. If I told you that hospitals were randomized to receive or not receive Program T, would that help fill in the remaining empty cells?

Randomization should make it so that hospitals that received Program T are the same, on average, as hospitals that did not receive Program T. Therefore, we would expect that same infant mortality in each of these groups had we made them all not use Program T, $\Pr(Y^{t=0}|T=0)=$

$\Pr(Y^{t=0}|T=1)$, and also if we had made them all use Program T, $\Pr(Y^{t=1}|T=0) = \Pr(Y^{t=1}|T=1)$). If we believe this, we can use information from hospitals observed to use Program T to say what would have happened to hospitals that did not use Program T had we made them use Program T. Similarly, we can use information from hospitals observed to not use Program T to say what would have happened to hospitals that did use Program T had we made them not use Program T. We call this assumption exchangeability and it allows us to say that, in each of the counterfactuals column of Table 3, we expect the infant mortality to be the same. If we believe the exchangeability assumption, we can calculate the values in the bottom row of the table and get our causal effect.

You might wonder what can be done if the hospitals were not randomized and there is confounding present. It may be that tertiary hospitals, which are more likely to see high-risk births, were much more likely to take up Program T than non-tertiary hospitals. As a result, we do not believe that the outcome in the hospitals without Program T is what would have happened in the hospitals with Program T had not used Program T. In other words, exchangeability is not satisfied. In this case, we must rely on conditional exchangeability instead. Conditional exchangeability simply assumes that exchangeability is met once we have conditioned on confounders. Another way to think of conditional exchangeability is that it assumes exchangeability within strata of confounders. Continuing with the example of tertiary hospitals as a confounder, we could assume that exchangeability is satisfied when looking only at tertiary hospitals or is satisfied only when looking at non-tertiary hospitals. We can carry out the above steps within tertiary and then non-tertiary hospitals, estimating the effect in each stratum and then pool the estimates. Conditional exchangeability is equivalent to assuming no confounding or selection bias after conditioning on or adjusting for confounders.

Table 3—By assuming exchangeability, we can fill in values for the bottom-left and top-right dark grey cells. Once the dark grey cells are filled in, we can take the average of the column to obtain the values in the bottom row. We can then contrast the values in the bottom row to get the causal effect.

Observed exposure (B)	Observed outcome (Y)	$Y^{a=0}$	$Y^{a=1}$
0	$\Pr(Y|A=0) = 7.6$	$\Pr(Y^{a=0}|A=0) = 7.6$	$\Pr(Y^{a=1}|A=0) = 6.9$
1	$\Pr(Y|A=1) = 6.9$	$\Pr(Y^{a=0}|A=1) = 7.6$	$\Pr(Y^{a=1}|A=1) = 6.9$
Total		$\Pr(Y^{a=0}) = 7.6$	$\Pr(Y^{a=1}) = 6.9$

Positivity

To illustrate positivity and how it can be violated, let us pick up from our analysis stratified on whether a hospital is tertiary. Imagine that we find that no non-tertiary hospitals used Program T (Table 4). If we tried to do what we mentioned previously of getting a causal estimate from each stratum and then pooling the estimates, we would find that we could not obtain an estimate among non-tertiary hospitals. Even if we put such data in a regression model, as is commonly done, it would return a causal effect of -1.7. This value, however, only applies to tertiary hospitals and not necessarily to all hospitals. This happens because the positivity assumption is violated. The positivity assumption is only satisfied when there are both treated and untreated observations in every stratum of confounders. If this is not the case, any estimate obtained may be a biased estimate of the true average causal effect.

The positivity assumption can be checked either by stratifying as above or, when there are many confounders, by using propensity scores. The following resources go into more depth on the positivity assumption.

Table 4—Repeating the above steps withing strata of whether the hospital is tertiary. In this example, positivity is not satisfied because there are no non-tertiary hospitals that used Program T.

Tertiary hospitals (H=1)

Observed exposure (T)	Observed outcome (Y)	$Y^{t=0}$	$Y^{t=1}$
0	$\Pr(Y\|T = 0, H = 1) = 8.6$	$\Pr(Y^{t=0}\|T = 0, H = 1) = 8.6$	$\Pr(Y^{t=1}\|T = 0, H = 1) = 6.9$
1	$\Pr(Y\|T = 1, H = 1) = 6.9$	$\Pr(Y^{t=0}\|T = 1, H = 1) = 8.6$	$\Pr(Y^{t=1}\|T = 1, H = 1) = 6.9$
Total		$\Pr(Y^{t=0}\|H = 1)$=8.6	$\Pr(Y^{t=1}\|H = 1) = 6.9$

Non-tertiary hospitals (H=0)

Observed exposure (T)	Observed outcome (Y)	$Y^{t=0}$	$Y^{t=1}$
0	$\Pr(Y\|T = 0, H = 0) = 6.6$	$\Pr(Y^{t=0}\|T = 0, H = 0) = 6.6$	$\Pr(Y^{t=1}\|T = 0, H = 0) = NA$
1	$\Pr(Y\|T = 1, H = 0) = NA$	$\Pr(Y^{t=0}\|T = 1, H = 0) = 6.6$	$\Pr(Y^{t=1}\|T = 1, H = 0) = NA$
Total		$\Pr(Y^{t=0}\|H = 0)$=6.6	$\Pr(Y^{t=1}\|H = 0) = NA$

Estimating causal effects

Filling in the tables as we have done is not necessary for a causal analysis. We have gone through these steps in order to give some intuition for what the causal assumptions are and how, when they are satisfied, an observed association could be interpreted as causal. Along with these causal assumptions, it must also be assumed that the variables in the analysis are not measured with error and that any model used in the analysis is correctly specified. Anytime one wishes to draw causal conclusions from the assumptions set discussed above, the plausibility of these assumptions must be discussed.

The assumption set outlined above is the most commonly in epidemiology to infer causation from data. There are many other assumption sets which, if satisfied, can also infer causation. In the next section we will discuss the assumptions required for instrumental variable analysis.

Causal assumptions for instrumental variable analysis

Let us return to our previous discussed scenario where our data come from a randomized trial of Program T but where the adherence to what hospitals were assigned to was not perfect. Some hospitals that were assigned to Program T were not able to implement it and some hospitals assigned to the control decided to implement Program T anyway. Can we still estimate the causal effect of all hospitals using Program T versus no hospitals using Program T (i.e., the per protocol effect)? Next we will present how we could use instrumental variable analysis to estimate this effect.

Intuition before assumptions

Think about how causation flows through a randomized trial. The randomization first affects the treatment. In a randomized trial with perfect adherence, randomization is a perfect determinant of treatment. When non-adherence is present, randomization affects treatment because people randomized to treatment are still more likely to be treated, but randomization does not determine treatment. After the randomization affects the treatment, the treatment then affects the outcome. Therefore, the effect of randomization on the outcome is just the effect of randomization on the treatment multiplied by the effect of treatment on the outcome:

$$\text{Total effect of Z on Y} = \text{Effect of Z on A} * \text{Effect of A on Y}$$

If we can estimate two pieces of this puzzle, i.e. the effect of the randomization on outcome and the effect of randomization on the treatment, we can use them to get the effect of the treatment on the outcome by rearranging the above equation as:

$$\text{Effect if A on Y} = \frac{\text{Total effect of Z on Y - Direct effect of Z on Y}}{\text{Effect of Z on A}}$$

In a randomized trial, we would expect that the group randomized to treatment and randomized to the control arm to be exchangeable due to the fact that it is randomized and therefore cannot be related to any confounders. This fact alone allows us to estimate the effect of randomization on both the treatment and the outcome. Essentially, all we have to do is divide the effect of randomization on outcome by the effect of randomization on the treatment.

We have played a bit fast and loose in this section in order to give you some intuition for how instrumental variable analysis works. We have also used the example of a randomized trial to do so. Instrumental variable analysis can, however, be also be used in observational data as well as along as the variable that is playing the role that randomization played in our example, meets the criteria of an instrumental variable which we will discuss now. As a reminder that the instrumental variable need not be randomization, I will refer to instrumental variables now with IV instead of Z.

Relevance

The relevance assumption is a simple one: the proposed instrumental variable must be associated with the treatment. This makes sense as we will be dividing by the association between the instrumental variable and the treatment. If the association is zero, then we will be dividing by zero. This can easily be checked by looking in the data to see if the instrumental variable is associated with the treatment. It should be noted that it is preferable for this association to be stronger rather than weaker to avoid an issue called weak instrument bias.

An astute reader will note that previously I was talking about the effect of randomization on the treatment and here I am using the word association instead. Though the intuition for instrumental variables is easier to explain when the instrumental variable causes the treatment, it turns out that the relationship between the instrumental variable and treatment does not need to be causal in nature (for reasons that are beyond our purposes here).

Exchangeability between the IV and Y

This assumption is identical to exchangeability in the previous section except now we are assuming exchangeability between IV and Y rather than between A and Y. In our randomized trial, where Z was random assignment, this exchangeability is expected through randomization. If we are considering other IVs, we will have to consider whether this assumption is satisfied.

A natural question is to wonder why we have bothered with instrumental variable analysis if it also relies on the assumption of exchangeability, the same assumption we previously said was very strong and we wish to avoid. The difference is we choose IVs where we have reason to believe that this assumption is plausible. Some examples of IVs where exchangeability is plausible is policy changes which affect treatment but are not related to risk factors for the outcome or genetic variants which can affect certain exposures.

Exclusion restriction

The last criteria is one we slipped into our intuitive example hoping the reader would not notice. We said that the effect of Z on the outcome is just the effect of Z on the treatment multiplied by the effect of the treatment on the outcome. This does not necessarily have to be the case. Z can also have an independent effect on the outcome that has nothing to do with treatment. For example, in an unblinded randomized trial of a vaccine, a person randomized to placebo may act differently, i.e. more cautious about potential infection, if they know they are randomized to placebo. This would be a case where Z has an effect on Y that is not due to the effect of the IV on the treatment. This assumption is called the exclusion restriction.

The fourth assumption: homogeneity or monotonicity

The three previous instrumental variable assumptions are what define an instrumental variable. We need one more assumption for the instrumental variable analysis and we can choose from two options.

If we believe that the effect of A on Y is the same in everyone, in other words that the effect is homogeneous, then we can interpret our causal effect as the average treatment effect which is just as it sounds: the average treatment effect in everyone. If we believe that the effect is not homogeneous, we can consider whether monotonicity is satisfied. Monotonicity exists when the relationship between Z and A goes in the same direction in everyone. In other words, if Z is possibility correlated with A in some people, there is not group of people in which Z and A are negatively correlated and vice versa. If monotonicity is satisfied but not homogeneity, the estimate is interpreted as a complier average effect. This means that the estimate only applies to people whose exposure is changed by the instrumental variable. For example, in the hospital example, compliers are hospitals who would implement Program T when randomized to do so and would not implement Program T when randomized not to do so. Hospitals that ignore randomization and would always implement Program T or never implement Program T would not contribute to the effect.

Conclusion

This has been whirlwind tour of two sets of causal assumptions.

Causal inference will always give answers of the sort, “if these assumptions hold, then the estimate may be interpreted as causal.” The assumptions of consistency and exchangeability cannot be checked in data though substantive knowledge can be used to argue for the degree to which the assumptions are likely to hold. In our example, maybe someone involved with the implement of Program T across our hypothetical might have knowledge of what considerations were involved in whether or not a hospital chose to implement Program T and therefore might be able to make strong arguments for which variables must be adjusted for for conditional exchangeability to hold.