

The Question of Causation

Disclaimer: This is a research area that spans statistics/epidemiology/philosophy and here you will only be given the very basics according to my study and understanding of the field

- Class exercise: Consider yourself omniscient. How would you prove that a change in X causes a change in Y. How, if you are omniscient would you be able to convince others of the causal effect. Think up an X and a Y and discuss with your neighbor (or give them aspirin X, headache Y, example).
 - **X causes Y**. In order to absolutely say whether a change in X **causes** a change in Y...you must know what Y is after X is changed **and** you must know what Y would have been if X had not been changed.
 - This is called knowing the “counterfactual”. The counterfactual is a fact that did not happen but would have happened if something else would have been changed. Maldonado and Greenland (2002) “Defining and estimating causal effects”, *International Journal of Epi*.
 - Another school of researchers describes this as “potential outcomes”. The potential outcomes are the results in Y that will occur under the different possible changes in X. Little and Rubin (2000) “Causal effects in Clinical and Epidemiological Studies via potential outcomes” *Annual Review of Public Health*.

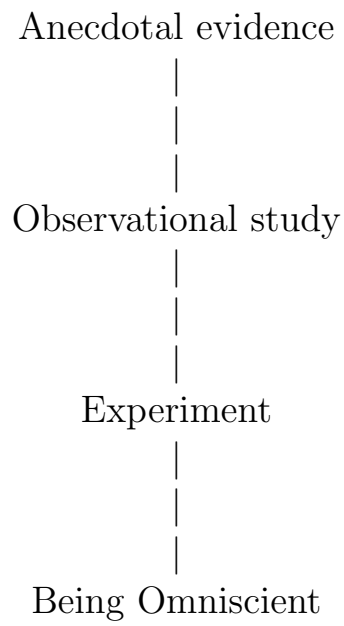
- Hypothetical Example (from Rubin’s short course on causality) Unit is you; Y is your assessment of your headache pain two hours after taking aspirin (action ASP) or not taking aspirin (action NOT)

unit	Y(ASP)	Y(NOT)	Causal effect
you	25	75	$Y(ASP)-Y(NOT) = -50$

- Of course no one is omniscient so how do we know that aspirin causes headache pain to decrease.
- Lots of past experience (“data”) related to headaches, there have been times in the past where you didn’t take an aspirin and you recall that your headache continued to hurt and there were times in past where you did take and aspirin and recall that your headache subsided. Use all that past data as a “proxy” or “substitute” for the potential outcomes $Y(ASP)$ and $Y(NOT)$ so that the causal effect can be estimated.

- The definition/description given above for X causes Y is based on a deterministic vision of the world. The deterministic world supposes that results during time $t + \Delta t$ will always be identical as long as the system going into time t is the same. This definition allows no room for chance (or “God”) to intervene.
- **X causes Y** in a stochastic world. In order to absolutely say whether a change in X **causes** a change in Y...you must know what the *probability distribution of Y* is after X has been changed **and** you must know what the *probability distribution of Y* would have been if X had not been changed.

Order of quality of evidence for estimating potential outcomes, i.e. quality of “proxy” or “substitute”.



Example of Anecdotal evidence: A person sees several commercials on tv that show patients at a particular cancer treatment facility X have successfully beaten cancer.

A possible causal hypothesis could be: if a person with cancer goes to cancer center X they will have a better chance of beating cancer than if that person would go to some other clinic.

What are being used as “proxies” to make this causal statement and how good are they?

- The potential outcome of $Y(\text{gotoclinic}X)$ is being estimated based on the smiling faces of successful patients in the commercials and the $Y(\text{gotosomeotherclinic})$ is being estimated with no data at all.
- Both proxies are poor and most definitely biased.

Example of an Observational study: A case control study was done to study the relationship between living near a power line and leukemia. M.S. Linet et al., “Residential exposure to magnetic fields and acute lymphoblastic leukemia in children”, *New England Journal of Medicine*, **337** (1997), pp. 1-7. A sample of 638 children who had leukemia and 620 who did not was taken. Magnetic-field levels at homes was recorded and broken into categories so that individuals were (for simplicity here) either exposed to power lines or not exposed to power lines.

A possible causal hypothesis could be: if a child lives in a home exposed to power lines he will have a greater chance of getting leukemia than if he lived in a home not exposed to power lines.

What is factual and what is counterfactual in the case control study in order to address the causal effect of power lines on leukemia?

- What is factual is that some kids are exposed to power lines and some **other** kids are not exposed to power lines.
- We observe the kids that are exposed to power lines and we can count how many get cancer, but we cannot observe what would have happened to these kids if they had not been exposed to power lines thus we have to use a “substitute” to measure this counterfactual.
- The kids that do not live near power lines are thus used as the “substitute” for the potential outcome of kids who do live near power lines.
- Comparing the odds of cancer in the factual group and the estimate of the odds in the substitute for the counterfactual we can get the causal effect.

The question then remains...how good is the estimate of the counterfactual, i.e. how valid is the substitute?

CONFOUNDING There can be other variables that are effecting both the X and the Y. In this example, it may be that family income is related to both living near power lines (X) and to getting leukemia (Y). If so, then the unexposed kids do not make a good substitute for the exposed kids. Observational studies usually try hard to consider possible confounding variables and then include them in the analysis as well.

Example of an Experiment (Clinical Trial): The Fen-Fen clinical trial by Guy-Grand, B., Apfelbaum, M, Crepaldi, G., Lefebvre, P. and Turner, P. “International trial of long term dexfenfluramine in obesity”, *Lancet*, **2**, 1142-1145 (1989) further discussed in Pocock, S. and Abdalla, M. “The hope and the hazards of using compliance data in randomized controlled trials”, *Statistics in Medicine*, 17, 303-317 (1998). 277 women randomized to dF and 310 randomized to placebo.

A possible causal hypothesis could be: The women in this study who take dF (“Fen-Fen”) will have a greater average decrease in their weight at 6 months than they would if they would have been the women in the study who did not take dF. What is factual and counterfactual for answering this causal statement using the data from the clinical trial?

- The factual is that the women in the study who took dF had some measurable weight loss. The counterfactual is what these same women would have seen if they hadn’t taken dF and instead would have taken the placebo.
- We will use the weight loss results of those other women who took the placebo as the “substitute” or estimate of potential outcome under not taking dF for the women who in fact took dF.
- Because the change in X here (i.e. treatment vs. placebo) was randomly assigned, we have theoretically a valid “substitute”. Those women who took dF should have results that are representative of what the other womens’ results would have been if they were the ones who got dF.
- No worry about confounders, since there is no extraneous variable that effects whether a woman got treatment or not.

Final note: Some act like an experiment provides definitive evidence about cause and effect, but experiments still have the problem of generalizability. You need to know who your target population is. Clinical trial participants are most often not a random sample of the population which will eventually be using/implementing the results.