Faculty of Health Sciences

SASA '23 Short Course on Causal Inference

Lecture 1 - Introduction to Causal Inference

What do we mean by Causal Inference

Much of the analysis of data in health and social sciences has as its central aim the quest to learn about cause-effect relationships.

- Does this treatment work?
- How harmful is the exposure?
- If we changed this hospital policy, would it reduce deaths in childbirth?

These are causal questions.

Randomised studies can answer such questions. But sometimes, we cannot do the randomisation, or we have data from an observational study. The goal of many observational studies is still to make inferences about the effects of causes, but in these settings, we need to care about confounding.

Example: HPV Vaccination

We want to see if the HPV vaccination impacts fertility. We have follow-up information about babies born to women who were old enough to participate in the 2014 school-based HPV vaccination program in the Western Cape region.
We have a cohort of 2000 women

- vaccination status in 2014 X,
- how many children they have had since vaccination Y
- Demographic and health-related measures at the time of vaccination $(Z_1, \ldots, Z_k)$
- Demographic and health-related measures in the following years since 2014, such as Cancer and death $(Q_1, \ldots, Q_k)$

Analysis

We look at the average number of babies born to vaccinated women $E\{Y|X=1\}$ and unvaccinated women $E\{Y|X=0\}$ and do a t-test, $E\{Y|X=1\} - E\{Y|X=0\}$.

We find a highly significant effect suggesting the unvaccinated women have more children and conclude that vaccination **causes** reduced fertility.

Not Causal!

Intuitively we know we can't give this a causal interpretation as there is likely to be confounding.

- Women receiving the HPV vaccine as girls are likely to have higher access to health care, and thus, birth control
- As it is school-based, women receiving the HPV vaccine as girls are more likely to be more highly educated
- there may be religious beliefs against both vaccination and birth control

So what have we estimated?

But What is it?

Probability and statistics let us describe aspects of the joint distribution of the variables observed in our dataset: means, variances, and regression parameters. However, this standard

language does not include a vocabulary for expressing how this distribution would change in response to an external intervention. What would the mean of births be if all girls that were eligible to

receive HPV vaccination did, versus if none did?

This is the causal question with a causal target **estimand**, but we can't even write that down in terms of sample means!

Counterfactual

So what is a counterfactual (or potential outcome). The idea is: Each subject in the target population has two potential outcomes:

- Y(X=0) or Y(0) the number of births if allocated to not have vaccination X = 0
- Y(X=1) or Y(1) the number of births if allocated to vaccination X = 1

At least one is counterfactual and will never be realized since women either were or were not vaccinated in 2014.

We want to target the average causal effect in the population, or the effect of vaccinating all eligible girls versus not vaccinating anyone, $E\{Y(1)\} - E\{Y(0)\}$

Confounding

When there is confounding, things are associated with the treatment/exposure and are potential causes of the outcome.

Then, the sample mean among those receiving vaccination $\frac{1}{\sum_i X_i} \sum_i Y_i * X_i$ is estimating $E\{Y|X=1\} \neq E\{Y(1)\}$

Consistency

If we observe someone under vaccination, this is the same value they would have had, had they been randomized to vaccination. $Y_i(x_i) = Y_i$ for a person i that is observed to receive x_i

Some people believe this is always true; some people believe you need to assume it. Either way, you need to know about it.

VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009 Nov;20(6):880-3. doi: 10.1097/EDE.0b013e3181bd5638. PMID: 19829187.

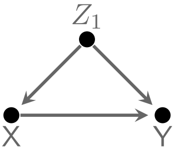
Direct Acyclic Graphs (DAG)

Components

- nodes (variables) vaccination, number of births, demographic variables
- edges, arrows the causal relationships, encoding a lack of independence

Confounders

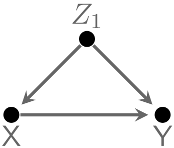
When there is a common cause of the exposure and the outcome there is confounding.



In this DAG Z_1 is a confounder of X and Y. If we used the observed sample means they would be biased for the causal effects.

Exchangeable

$Y(X) \perp\!\!\!\perp X$ then there is no confounding



$Y(X) \perp\!\!\!\perp X|Z_1$ then there is no unmeasured confounding

In both cases, we have Exchangeability; in the second case, after conditioning on Z_1. In both cases we have a means of estimating the causal effect we are interested in.

Case 1

Why does exchangeability mean that we can estimate a causal effect?

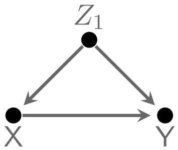
This holds, because

$E\{Y(1)\} = E\{Y(1)|X = 1\}$ by exchangeability, and

$E\{Y(1)|X = 1\} = E\{Y|X = 1\}$ by consistency.

This is why randomized studies are so important!

Case 2



We can estimate this because: $E\{Y(1)\} \neq E\{Y(1)|X=1\}$, but $E\{Y(1)|Z_1\} = E\{Y(1)|X=1, Z_1\}$ by exchangeability, and $E\{Y(1)|X=1, Z_1\} = E\{Y|X=1, Z_1\}$ by consistency.

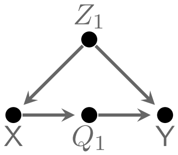
How do we know what are confounders?

We either do or we do not, there isn't a way to tell from the data directly.

Why don't we just adjust for everything?

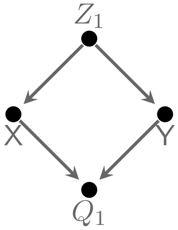
Adjusting for everything is not a good idea, because there are more things than confounders!

Mediators



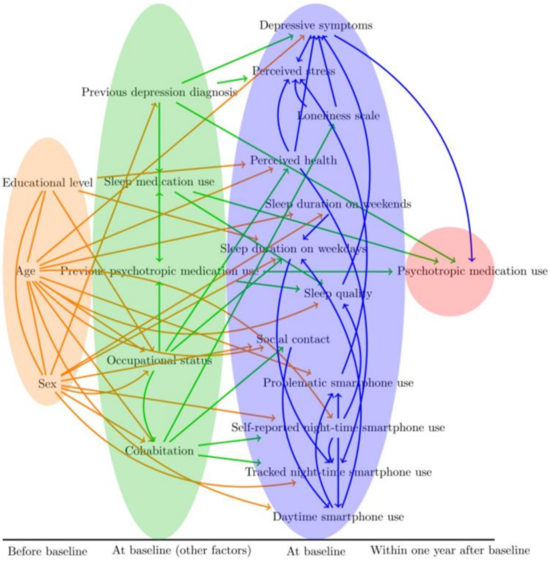
Q_1 here is a mediator (a complete mediator); it is along the causal pathway from X to Y, adjusting for it and Z_1 will remove any association between X and Y. There is a causal effect of X on Y here, but only via Q_1. For example, the HPV vaccination on fertility conditioning on HPV infection status after vaccination.

Colliders



Here there is no causal effect of X on Y; you must condition on Z_1, to estimate this lack of an effect, but if you also condition on Q_1 you will see a statistical association because you have opened a pathway from X to Y.

Examples from real life



What should I condition on?

There exists a proven algorithm for determining a sufficient adjustment set, i.e. a set of confounders conditioning on which provides conditional exchangeability.

You don't need to know that algorithm, it is based on Pearl's do-calculus. But, you generally don't need to implement it yourself. The *Dagitty* program, which we will use in the exercise, will help!