Causal Inference & Causal Learning

Confounding & Selection bias

Xiaolei Lin

School of Data Science
Fudan University

April 12, 2024

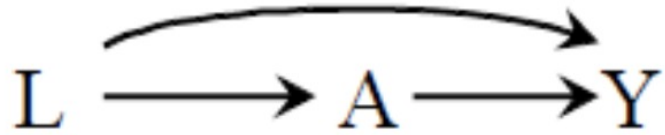
Recap

- g-formula in Randomized experiments and Observational studies
- Effect modification & interaction
- DAG & d-separation

Today's plan

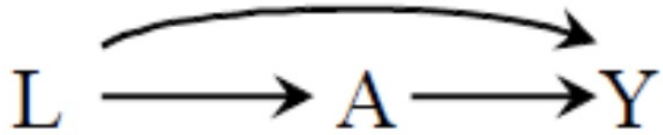
- back-door criterion & confouding
- collider & selection bias
- methods to adjust for both types of bias

Confounding & Backdoor criterion



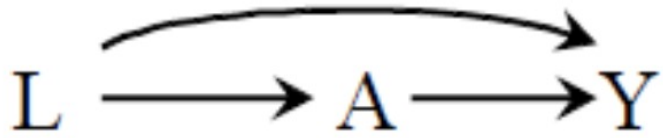
- Confounding bias arises when the treatment and outcome share a common cause
- The DAG shows two sources of association
  - the path $A \rightarrow Y$ that represents the causal effect of A on Y
  - the path $A \leftarrow L \rightarrow Y$ between A and Y that includes the common cause L
- The path $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is called a **backdoor path**

Confounding & Backdoor criterion



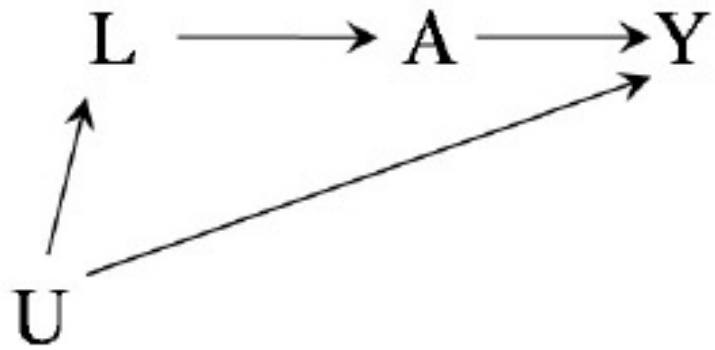
- Confounding bias arises when the treatment and outcome share a common cause
- because of confounding, the associational risk ratio does $\frac{Pr(Y=1|A=1)}{Pr(Y=1|A=0)}$ not equal the causal risk ratio $\frac{Pr(Y^{a=1}=1)}{Pr(Y^{a=0}=1)}$
- when there is confounding, association is not causation

Confounding in observational studies



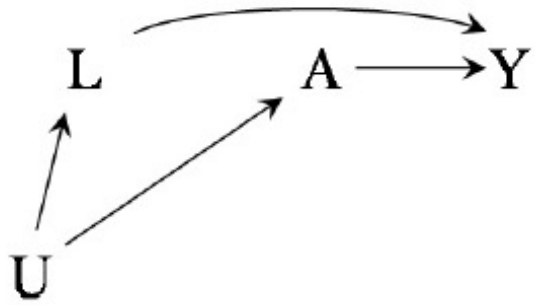
- healthy worker bias: The effect of working as a firefighter A on the risk of death Y will be confounded if "being physically fit" L is a cause of both being an active firefighter and having a lower mortality risk.

Confounding in observational studies



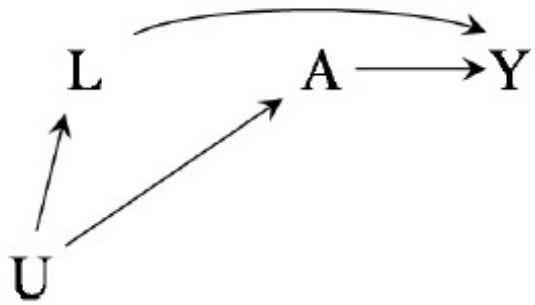
- Clinical decisions: The effect of aspirin A on the risk of stroke Y will be confounded if aspirin is more likely to be prescribed to individuals with heart disease L, which is both an indication for aspirin and a risk factor for stroke.
- both L and Y are caused by atherosclerosis U, an unmeasured variable

Confounding in observational studies



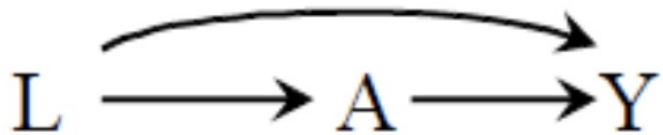
- Lifestyle: The effect of lack of exercise A on the risk of death Y will be confounded if lack of exercise is associated with another behavior L (cigarette smoking) that has a causal effect on Y and tends to co-occur with A
- the unmeasured variable U represents personality and social factors that lead to both lack of exercise and smoking
- reverse causation: subclinical disease U results both in lack of exercise A and an increased risk of clinical disease Y, when L is unknown

Confounding in observational studies



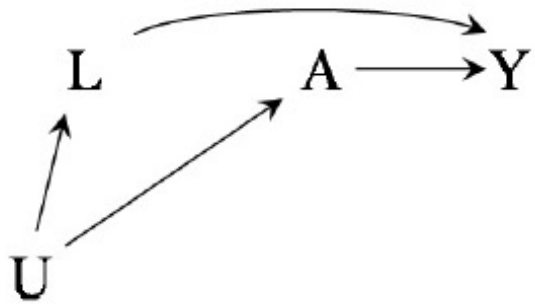
- Genetic factors: The effect of a DNA sequence A on the risk of developing certain trait Y will be confounded if there exists a DNA sequence L that has a causal effect on Y and is more frequent among people carrying A
- linkage disequilibrium or population stratification
- U: ethnicity or other factors that result in linkage of DNA sequences

Confounding in observational studies



- Social factors: The effect of income at age 65 A on the level of disability at age 75 Y will be confounded if the level of disability at age 55 L affects both future income and disability level.

Confounding in observational studies



- Environmental exposures: The effect of airborne particulate matter A on the risk of coronary heart disease Y will be confounded if other pollutants L whose levels co-vary with those of A cause coronary heart disease
- unmeasured variable U: weather conditions that affect the levels of all types of air pollution

Confounding in observational studies

In all above cases, the bias has the same structure:

- it is due to the presence of a common cause (L or U) that is shared by the treatment A and the outcome Y , which results in an open backdoor path between A and Y

Confounding & exchangeability

- In practice, if we believe confounding is likely, a key question arises: can we determine whether there exists a set of measured covariates L for which conditional exchangeability holds?
- answering this question is possible if one knows the causal DAG that generated the data
- A set of covariates L satisfies thebackdoor criterion if all backdoor paths between A and Y are blocked by conditioning on L and L contains no variables that are descendants of treatment A.
- conditional exchangeability $Y^a \perp A \mid L$ holds if and only if L satisfies the backdoor criterion

Backdoor criterion

- Randomization eliminates all backdoor paths by severing the association between A and L
- More generally, in the absence of randomization, the backdoor criterion states that the treatment effect is identified if one has observed enough variables to block all backdoor paths, that is if treatment and outcome are d-separated given the measured covariates in a graph in which the arrow out of treatment are removed

Backdoor criterion

This criterion answers three questions:

1. does confounding exist?
2. can confounding be eliminated?
3. what variables are necessary to eliminate the confounding?

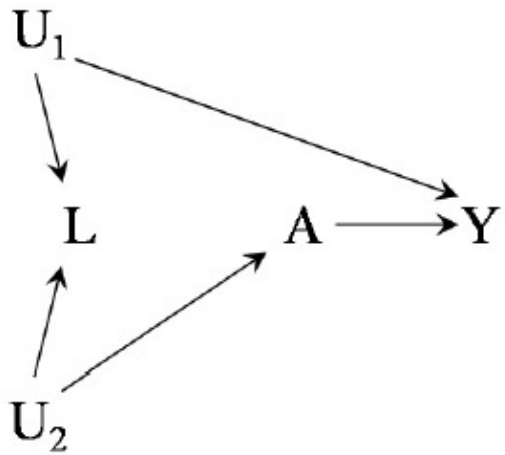
Crucially it can be used to decide whether one has measured a sufficient set of "confounders" to block all backdoor paths and therefore to adjust for confounding

Backdoor criterion

This criterion does not answer:

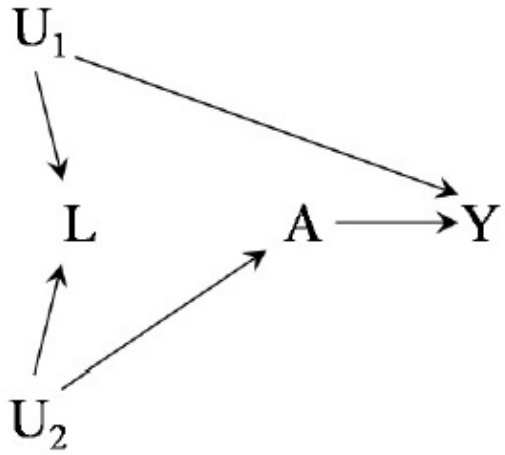
1. the mgnitude of confounding.
   - Some unblocked backdoor paths are weak (e.g., if L does not have a large effect on either A or Y) and thus induce little bias
2. direction of confounding
   - several strong backdoor paths induce bias in opposite directions and thus result in a weak net bias

Backdoor criterion



Is there confounding between A and Y?

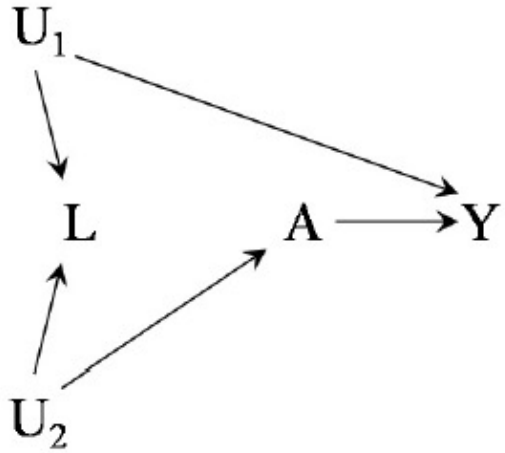
Backdoor criterion



- The backdoor path between A and Y through L ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$) is blocked because L is a collider on that path.

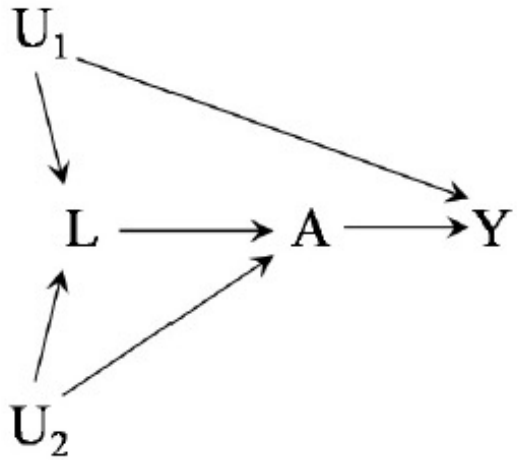
- A: physical activity, Y: cervical cancer, L: diagnostic test (Pap smear), U_1: precancer lesion, , and U_2: health conscious personality (more physically active, more visits to the doctor).

Backdoor criterion



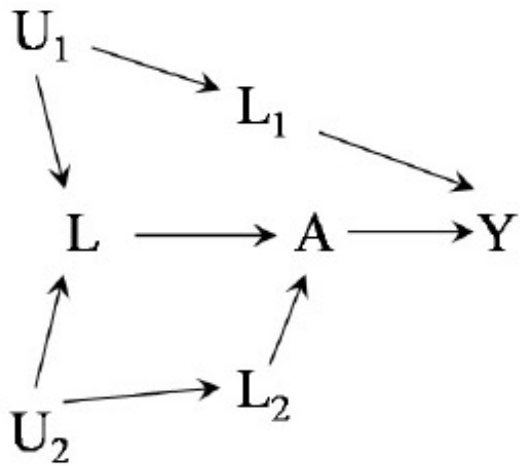
- conditional exchangeability given L does not hold
- counterfactual risks $Pr[Y^{a=1} \mid L = l]$ are not equal to the stratum-specific risks $Pr[Y = 1 \mid A = a, L = l]$
- adjustment for L via standardization $\sum_l Pr[Y = 1 \mid A = a, L = l] \times Pr[L = l]$ gives a biased estimate of $Pr[Y^a]$.
- adjustment for L would induce bias because conditioning on the collider L opens the backdoor path between A and Y $(A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y)$

Backdoor criterion



- $L \rightarrow A$ creates an open backdoor path $A \leftarrow L \leftarrow U_1 \rightarrow Y$ because U_1 is a common cause of A and Y , and so confounding exists.
- conditional on L blocks the backdoor path $A \leftarrow L \leftarrow U_1 \rightarrow Y$, but opens a backdoor path on which L is a collider $(A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y)$
- bias is intractable

Backdoor criterion



solution:

- measure a variable L_1 between U_1 and either A or Y, such that conditional exchangeability given L_1
- measure a variable L_2 between U_2 and either A or L, such that conditional exchangeability given L_2 and L

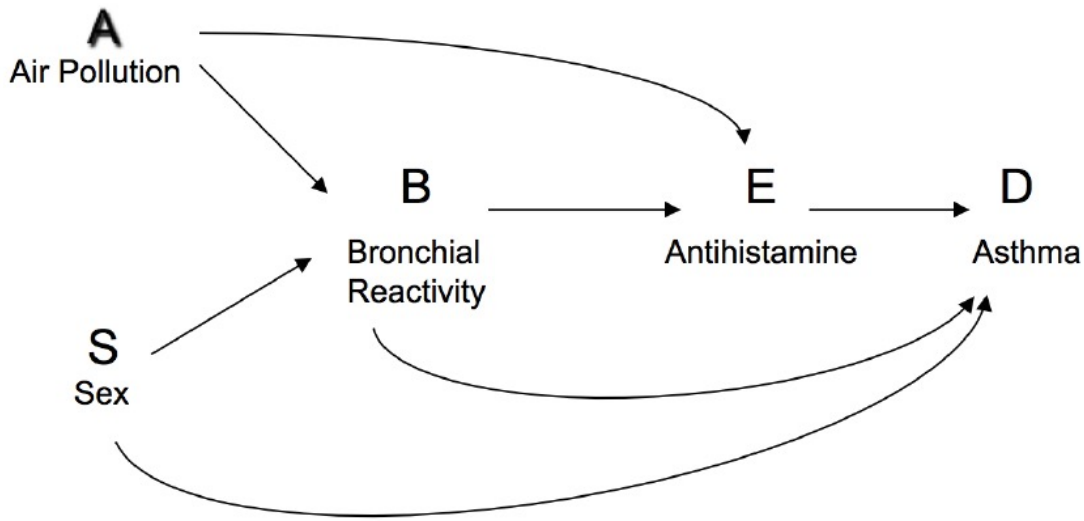
Backdoor criterion

A confounder was traditionally defined as any variable that meets the following three conditions:

1. It is associated with the treatment
2. It is associated with the outcome conditional on the treatment
3. It does not lie on a causal pathway between treatment and outcome

A collider variable meets the same conditions!

Directed Acyclic Graphs (DAGs)

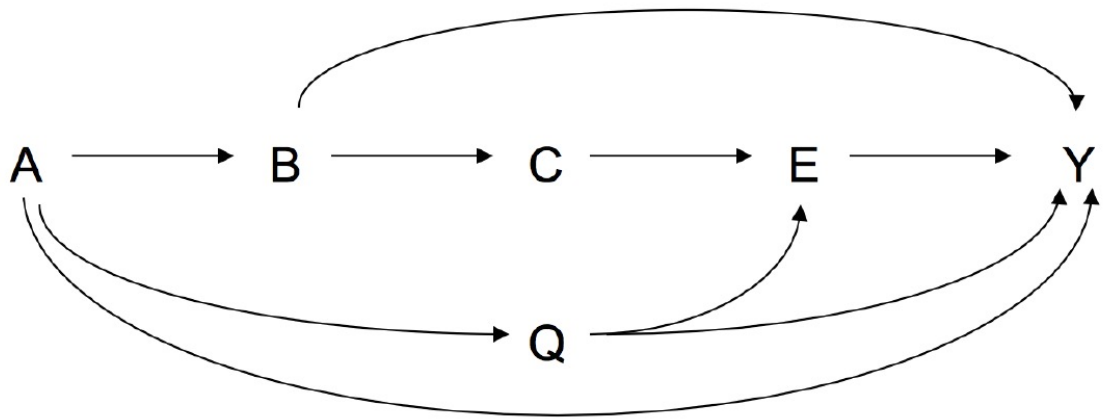


A causal DAG is defined by non-parametric structural equations: $V = f_V(pa_V, \epsilon_V)$ with all V independent (Pearl, 1995).

- $D = f_D(E, B, S, \epsilon_D)$, $E = f_E(A, B, \epsilon_E)$, $B = f_B(A, S, \epsilon_B)$, $A = f_A(\epsilon_A)$, $S = f_S(\epsilon_S)$

Case study: Backdoor Path Criterion

Backdoor Path Criterion (Pearl 1995): For exposure E and outcome Y , if a set of variables Z is such that no variable in Z is a descendent of E and Z blocks all "back-door paths" from E to Y (i.e. all paths from E to Y with edges into E) then conditioning on Z suffices to control for confounding.

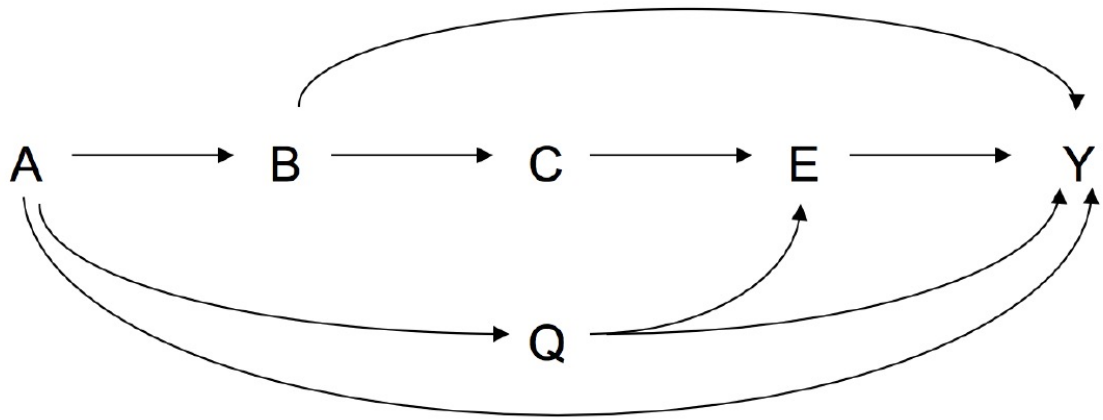


For the causal effect of E on Y :

- Controlling for C and Q suffices?
- Controlling for B and Q suffices?
- Controlling for A and Q suffices?

Case study: Backdoor Path Criterion

Backdoor Path Criterion (Pearl 1995): For exposure E and outcome Y , if a set of variables Z is such that no variable in Z is a descendent of E and Z blocks all "back-door paths" from E to Y (i.e. all paths from E to Y with edges into E) then conditioning on Z suffices to control for confounding.

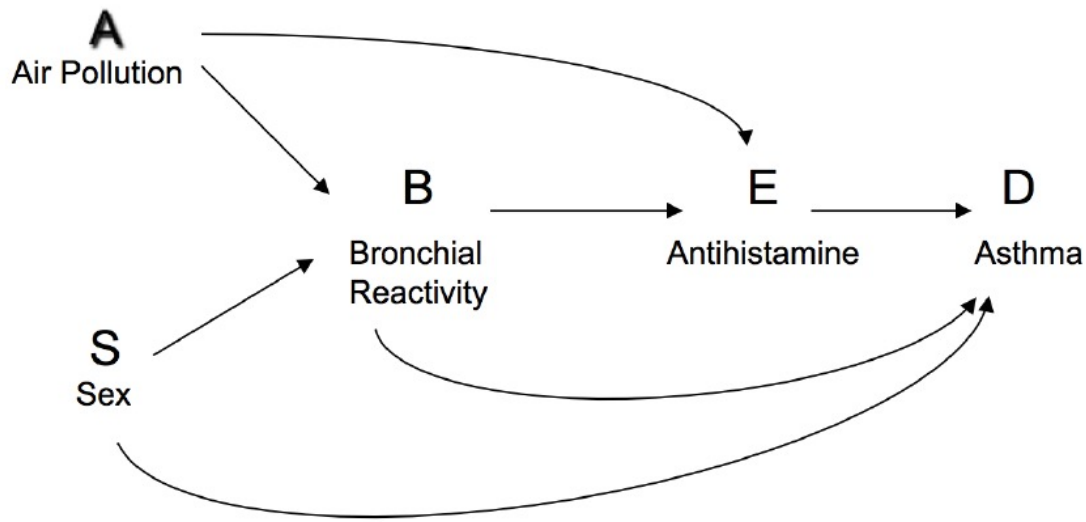


For the causal effect of E on Y :

- Controlling for C and Q suffices
- Controlling for B and Q suffices
- Controlling for A and Q does NOT suffice ($E - C - B - Y$ unblocked)

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma

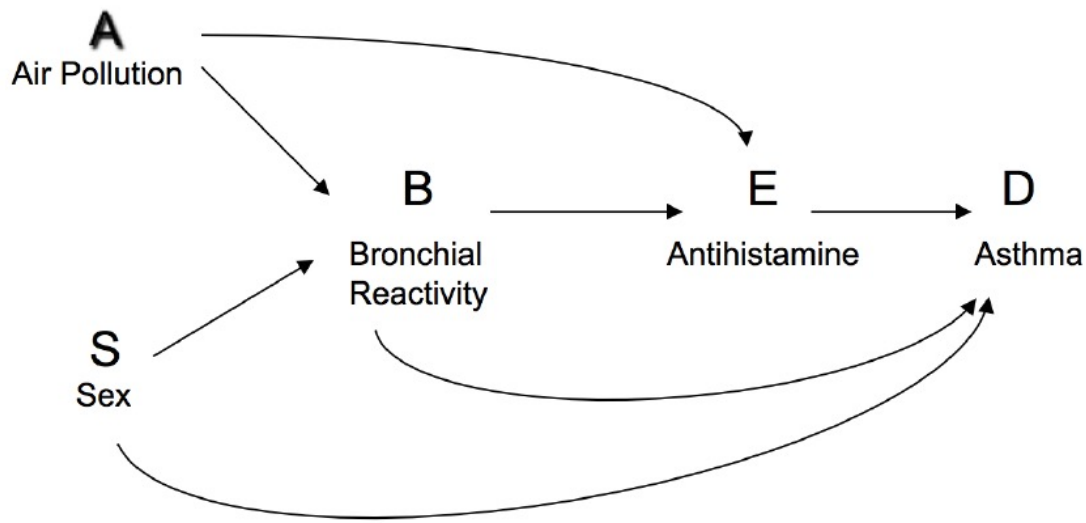


Assumptions in the DAG:

- A affects D only through B and E
- S affects E only through B
- There are no common causes of two variables of the graph that are not on the graph

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma

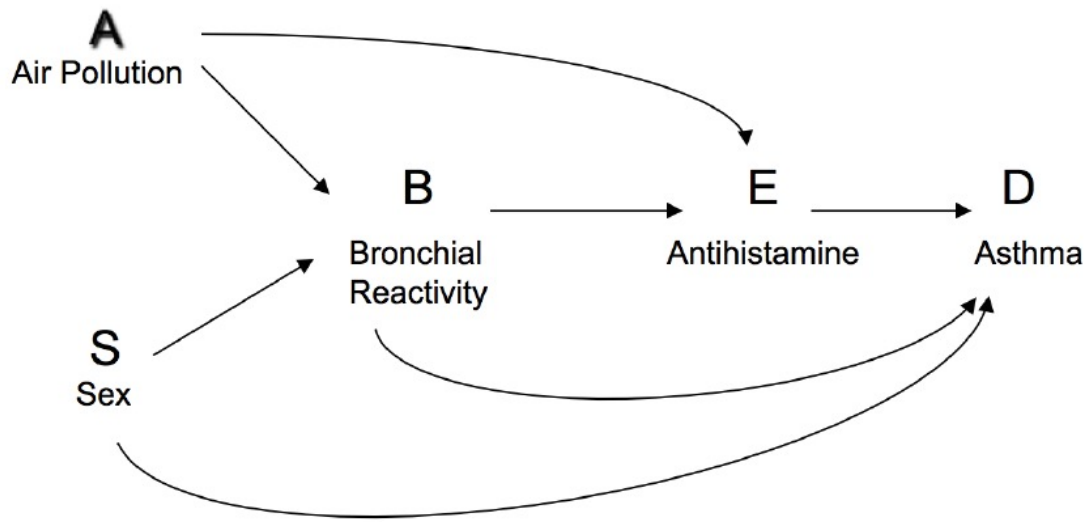


To control for confounding

- condition on A, B and S suffice?
- condition on only A and B suffice?
- condition on only S and B suffice?
- condition on S alone suffice?

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma



To control for confounding

- it suffices to condition on A, B and S
- it suffices to condition on only A and B
- it suffices to condition on only S and B
- it does not suffice to condition on S alone

Regression and Causation

Let Y denote a continuous outcome, A exposure and C covariates

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

- If the linear regression is correctly specified, but C does not include all the confounders, regression coefficients do not have a causal interpretation but do have an associational interpretation:
- "If we randomly select two individuals from a population and both have the same value of C but the second individual has a value of A one unit higher than the first, then on average, the second individual will have a value of Y that is β_1 units higher."
- Many research studies will appropriately qualify their findings, noting that their results concern association amongst variables and do not necessarily imply causal relationships

Regression and Causation

Regression and Causation: For regression coefficients to have a causal interpretation we need both that

1. The linear regression to be correctly specified
2. All confounders of, e.g., the relationship between treatment A and Y be in the model.

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

if $Y^a \perp A \mid C$,

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = E[Y = 1 \mid A = 1, C = c] - E[Y = 0 \mid A$$

i.e., intervening to increase A by one unit will, on average, increase Y by β_1 units.

Regression and Causation

In the absence of interactions between A and C in linear regression:

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

if $Y^a \perp A \mid C$,

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = E[Y = 1 \mid A = 1, C = c] - E[Y = 0 \mid A$$

Regression and Causation

In the presence of interactions between A and C in linear regression:

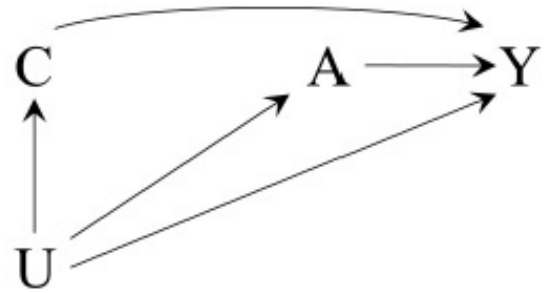
$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C + \beta_3 A \times C$$

For conditional average causal effects:

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = \beta_1 + \beta_3 \times C$$

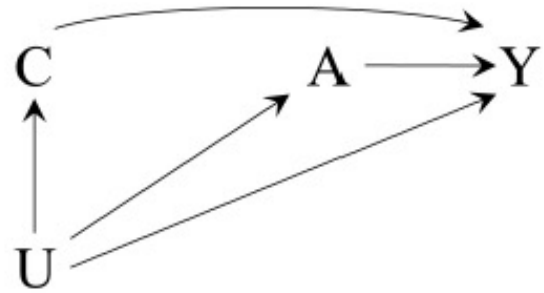
What are the implications for inference? The standard error of the effect needs to account for the uncertainty in the estimation of the covariate distributions! How can we account for it? (bootstrap!)

Negative control and Causation



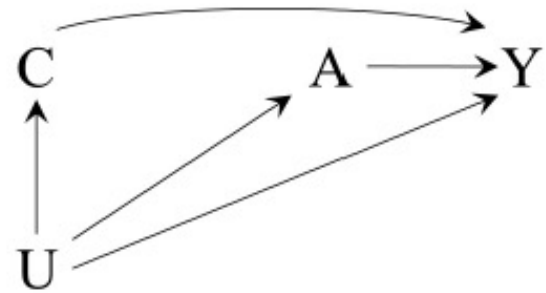
- Goal: compute average causal effect of aspirin A (1: yes; no: no) on blood pressure Y
- unmeasured common causes U of A and Y, such as history of heart disease
- IP weighting and g-formula cannot work due to unmeasured confounding

Negative control and Causation



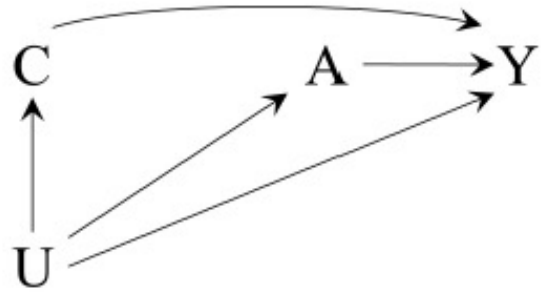
- Goal: compute average causal effect of aspirin A (1: yes; no: no) on blood pressure Y
- unmeasured common causes U of A and Y, such as history of heart disease
- IP weighting and g-formula cannot work due to unmeasured confounding

Negative control and Causation



- C: pre-treatment blood pressure, measured right before the treatment
- the causal effect of A on C is 0
- $E(C \mid A = 1) - E(C \mid A = 0) \neq 0$ measures the magnitude of confounding for the effect of A on C on the additive scale

Negative control and Causation



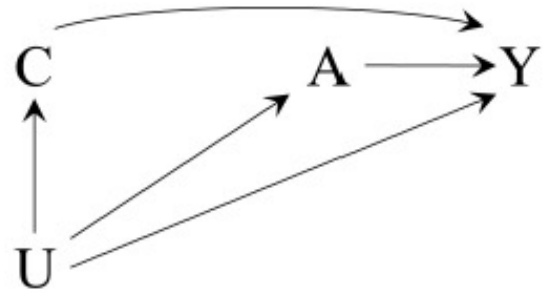
- If the magnitude of additive confounding for the effect of A on the negative outcome control C is the same as for the effect of A on outcome Y , then we can compute the effect of A on Y in the treated
- under additive equi-confounding

$$E(Y^0 \mid A = 1) - E(Y^0 \mid A = 0) = E(C \mid A = 1) - E(C \mid A = 0)$$

- the causal effect of A on Y

$$E(Y^1 - Y^0 \mid A = 1) = (E(Y \mid A = 1) - E(Y \mid A = 0)) - (E(C \mid A =$$

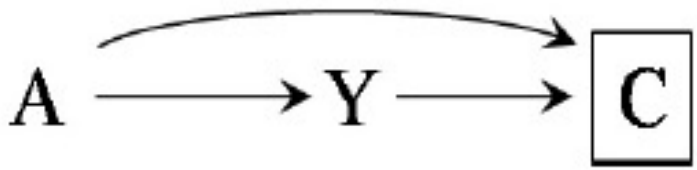
Negative control and Causation



- This method for confounding adjustment is known as difference-in-differences
- it requires measurement of the outcome both pre- and post-treatment (or at least that the true outcome Y and the C are measured on the same scale)
- it requires additive equi-confounding

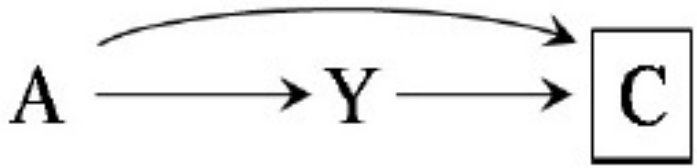
Structure of selection bias

selection bias: various biases that arise from the procedure by which individuals are selected into the analysis.



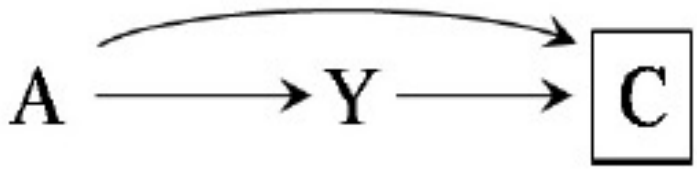
- A: folic acid supplements given to pregnant women shortly after conception, binary
- Y: cardiac malformation (1: yes, 0: no)
- C: death before birth (1: death, 0: alive)

Structure of selection bias



- A cardiac malformation increases mortality (arrow from Y to C)
- folic acid supplementation decreases mortality by reducing the risk of malformations other than cardiac ones (arrow from A to C)
- but study was restricted to fetuses who survived until birth ($C = 0$)

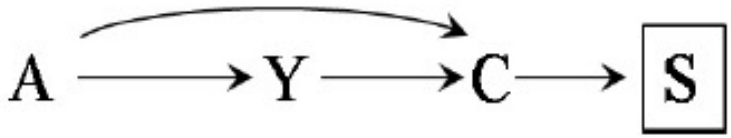
Structure of selection bias



two sources of association between treatment and outcome

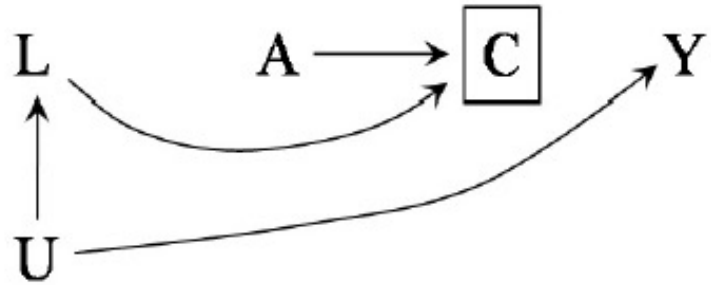
- the open path $A \rightarrow Y$ that represents the causal effect of A on Y
- the open path $A \rightarrow C \leftarrow Y$ that links A and Y through their (conditioned on) common effect C
- induced association between the treatment A and the outcome Y due to conditioning on C as selection bias

Structure of selection bias



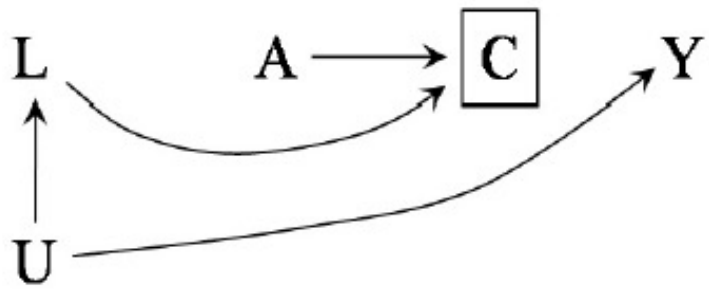
- S: parental grief (1: yes, 0: no), affected by vital status at birth
- the study was restricted to non grieving parents $S = 0$ because the others were unwilling to participate
- conditioning on a variable S affected by the collider C also opens the path $A \rightarrow C \leftarrow Y$

Structure of selection bias



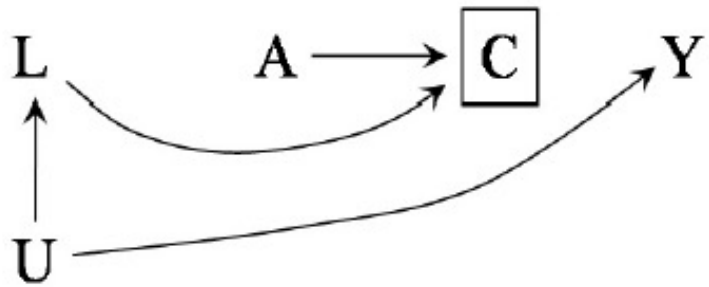
- A: antiretroviral treatment
- Y: death after 3 year follow-up
- U: high level of immunosuppression (1: yes, 0: no)
- C: censor status (Individuals who drop out from the study or are otherwise lost to follow-up)
- L: presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma (unmeasured)

Structure of selection bias



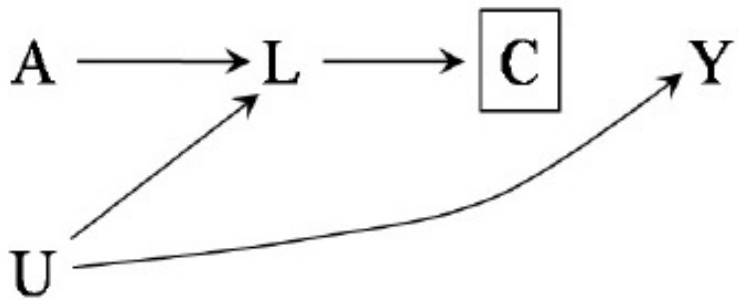
- $A \rightarrow C$: Individuals receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout
- $U \rightarrow L \rightarrow C$: high level of immunosupression increase the probability of censor via side effects L
- $U \rightarrow Y$: high level of immunosuppression increase 3-year mortality
- analysis was restricted to individuals who remained uncensored $C = 0$ because those are the only ones in which Y can be assessed

Structure of selection bias



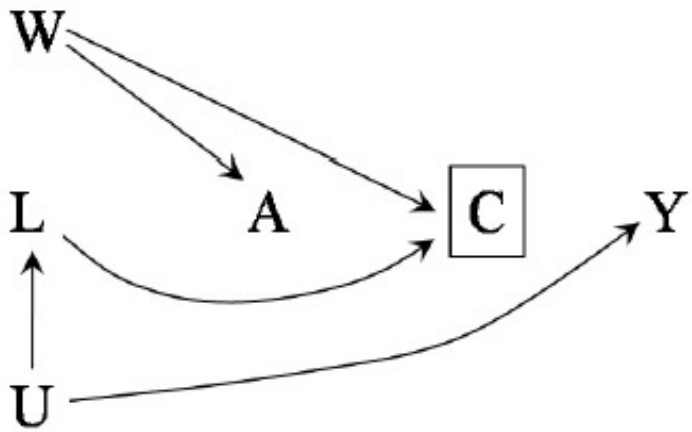
- conditioning on the collider C opens the path $A \rightarrow C \leftarrow L \leftarrow U \rightarrow Y$
- C is common effect of treatment A and of a cause U of the outcome Y, rather than a common effect of treatment and outcome

Structure of selection bias



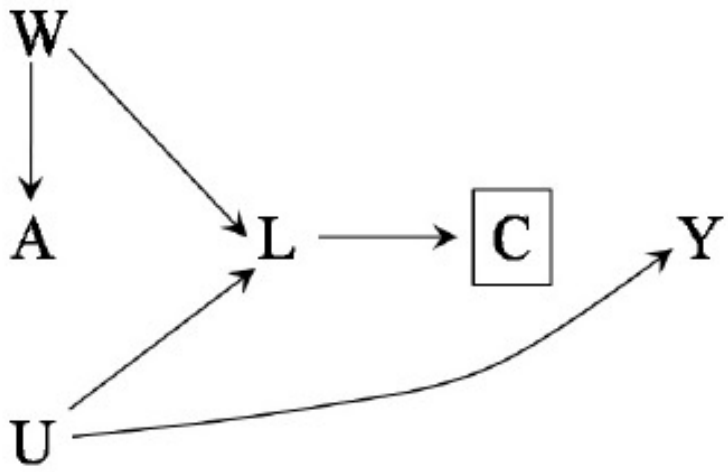
- treatment A has a direct effect on symptoms L
- restricting the study to uncensored individuals implies conditioning on common effect C of A and U, thus introducing selection bias

Structure of selection bias



- W: unmeasured lifestyle/personality/educational variables
- W determine both treatment (arrow from W to A) and attitudes toward attending study visits (arrow from W to C)

Structure of selection bias



- W: unmeasured lifestyle/personality/educational variables
- W determine both treatment (arrow from W to A) and threshold for reporting symptoms (arrow from W to L).

Examples of selection bias

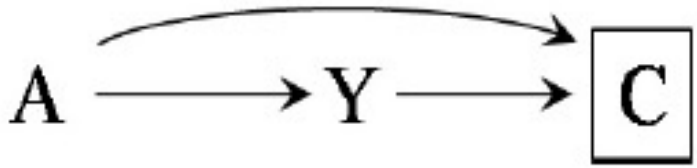
- differential loss to follow-up: also referred to as informative censoring
- missing data bias (nonresponse bias): C represent missing data for any reason, not just as a results of loss to follow-up.
- healthy worker bias: occupational chemical exposure A on mortality Y in a cohort of factory workers. The underlying unmeasured true health status U is a determinant of both death Y and of being at work C (1: no, 0: yes). The study is restricted to individuals who are at work ($C = 0$) at the time of outcome ascertainment

Examples of selection bias

- self selection bias (volunteer bias): A is cigarette smoking, Y is coronary heart disease, U is family history of heart disease, and $W$is healthy lifestyle. Under any of these structures, selection bias may be present if the study is restricted to those who volunteered or elected to participate ($C = 0$).

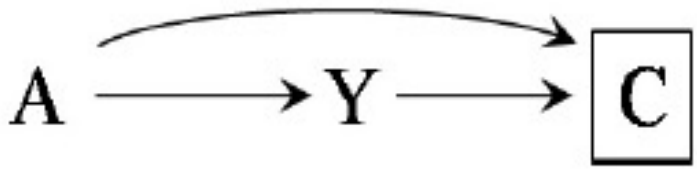
- selection affected by treatment received before study entry: C represents selection into the study (1: no, 0: yes) and that treatment A took place before the study started. If treatment affects the probability of being selected into the study, then selection bias is expected (generalization of self selection)

Selection bias in case-control study



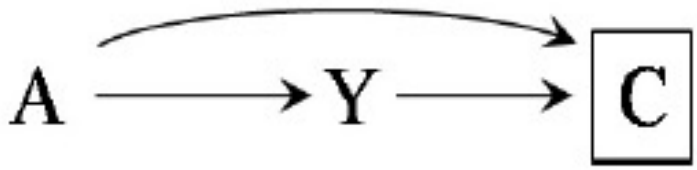
- A: postmenopausal estrogen treatment
- Y: coronary heart disease
- C: whether the women in the study population is selected for the case control study

Selection bias in case-control study



- In this particular case-control study, the investigator decided to select controls ($Y = 0$) preferentially among women with a hip fracture
- treatment A has a protective causal effect on hip fracture
- the selection of controls with hip fracture implies that treatment A now has a causal effect on selection C

Selection bias in case-control study

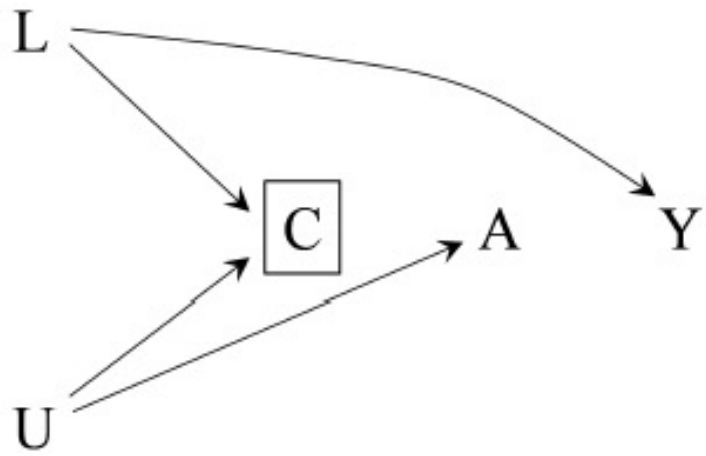


- If individuals with hip fracture are oversampled as controls, then the probability of control selection depends on a consequence of treatment A (as represented by the path from A to C)
- this bias arises because we are conditioning on a common effect C of treatment and outcome
- intuitive explanation: among individuals selected for the study, controls are more likely than cases to have hip fracture. Therefore, because estrogens lower the incidence of hip fractures, a control is less likely to be on estrogens than a case, and hence the $A - Y$ odds ratio conditional on C would be greater than the causal odds ratio in the population

Selection & confounding

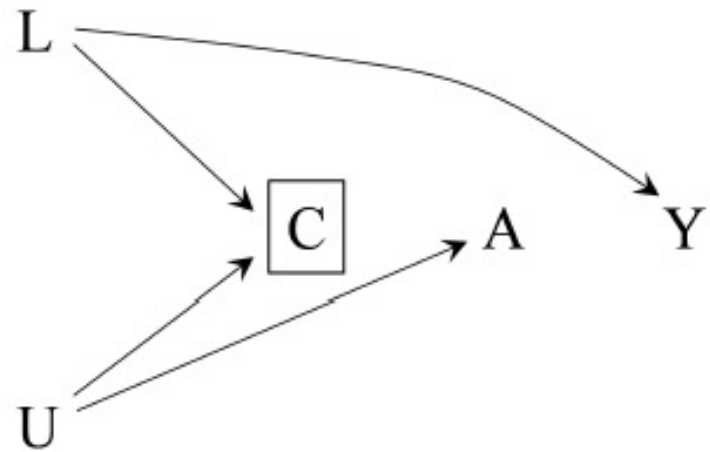
- A key difference between confounding and selection bias: randomization protects against confounding, but not against selection bias when the selection occurs after the randomization.
- some disciplines use confounding and selection bias interchangably: social science often refore to unmeasured confounding as selection on unobservables, econometricians often use the term "selection bias" to refer to both types of biases
- both bias result in lack of exchangeability between the treated and controls

Selection & confounding



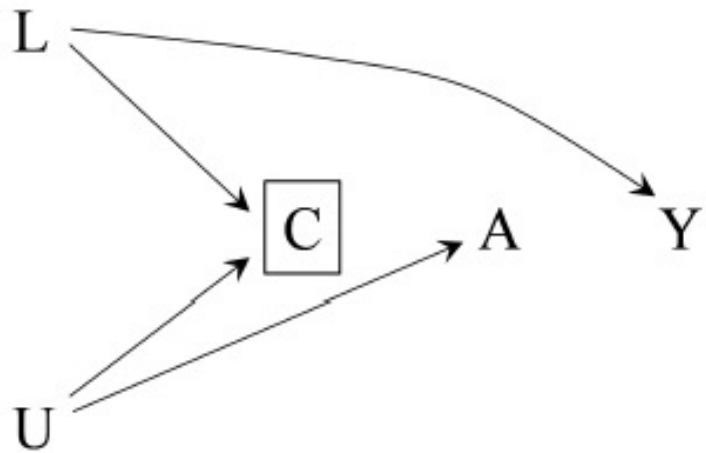
- A: being physically active
- Y: heart disease
- L: parental socioeconomic status
- C: being a firefighter
- U: attraction toward activities (involving physical activity)

Selection & confounding



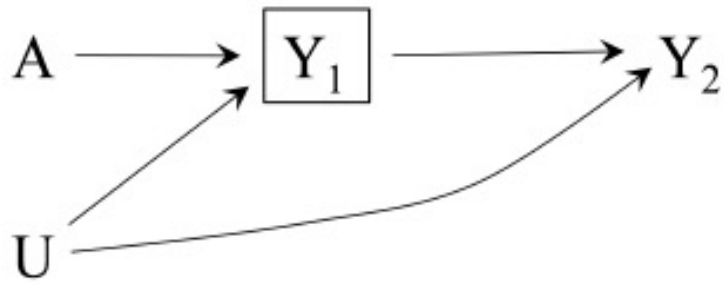
- is there confounding unconditional on C?

Selection & confounding



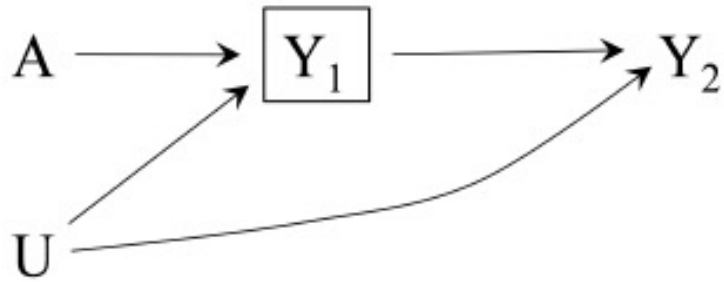
- is there confounding if restricted to firefighters?
- is it necessary to adjust for L in studies not restricted to firefighters?

The built-in selection bias of hazard ratio



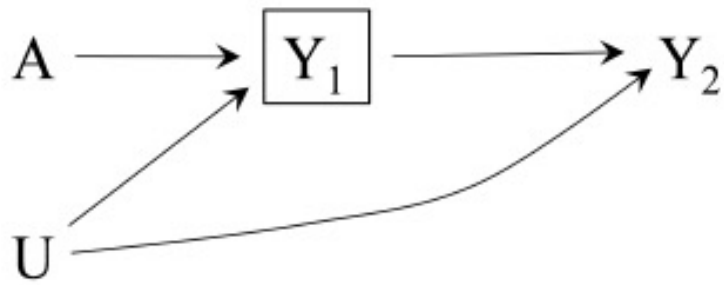
- A: heart transplant
- Y_1: death at time 1 (heart transplant decreases the risk of death at time 1)
- Y_2: death at time 2 (heart transplant has no direct effect on death at time 2)
- U: unmeasured haplotype that decreases the individual's risk of death at all times

The built-in selection bias of hazard ratio



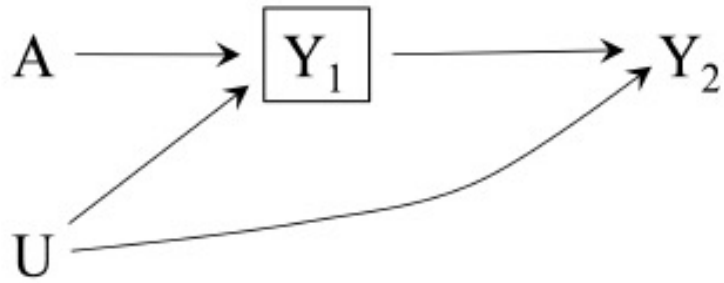
- no confounding between A and Y_1
- associational risk ratio $aRR_{AY_1} = \frac{Pr(Y_1=1|A=1)}{Pr(Y_1=1|A=0)}$ and $aRR_{AY_2} = \frac{Pr(Y_2=1|A=1)}{Pr(Y_2=1|A=0)}$ are unbiased estimator of the causal effect of A on Y_1 and Y_2

The built-in selection bias of hazard ratio



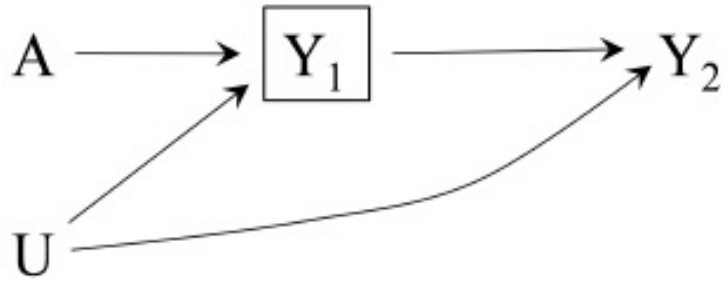
- hazard of A on Y_1: probability of dying at time 1
- hazard of A on Y_2: probability of dying at time 2 among those who survived past time 1 $aHR_{AY_2|Y_1=0} = \frac{Pr(Y_2=1,A=1,Y_1=0)}{Pr(Y_2=1,A=0,Y_1=0)}$
- those who are treated and survive at time 1 are less likely than those who are untreated but survived at time 1 to have the protective haplotype U, and therefore are more likely to die at time 2

The built-in selection bias of hazard ratio



- conditional on Y_1, heart transplant A is associated with a higher mortality at time 2, compared to at time 1
- hazard ratio at time 2 is a biased estimate of the direct effect of treatment on mortality at time 2, due to selection bias from open backdoor path $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ between A and Y_2
- in survival analysis literature, an unmeasured cause of death that is marginally unassociated with treatment such as U is often referred to as a **frailty**

The built-in selection bias of hazard ratio



- the conditional hazard ratio $aHR_{AY_2} \mid Y_1 = 0, U = 1$ within each stratum of U because the path $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ is now blocked by conditioning on the non-collider U
- the unconditional hazard ratio $aHR_{AY_2} \mid Y_1 = 0$ differs from the stratum specific hazard ratios $aHR_{AY_2} \mid Y_1 = 0, U$, even though U is independent of A (noncollapsibility of the hazard ratio)
- In the absence of data on U, it is impossible to know whether A has a direct effect on Y_2

Selection & censoring

An investigator conducted a marginally randomized experiment to estimate the average causal effect of wasabi intake on the one-year risk of death ($Y = 1$).

Half of the 60 study participants were randomly assigned to eating meals supplemented with wasabi ($A = 1$) until the end of follow-up or death, whichever occurred first.

The other half were assigned to meals that contained no wasabi ($A = 0$). After 1 year, 17 individuals died in each group.

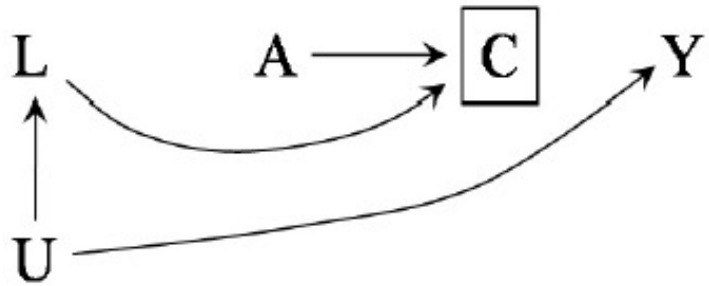
Selection & censoring

- The associational risk ratio
  $Pr(Y = 1 \mid A = 1)/Pr(Y = 1 \mid A = 10 = 1$
- Because of randomization, the causal risk ratio
  $Pr(Y^{a=1} = 1)/Pr(Y^{a=0} = 1) = 1$
- Unfortunately, the investigator could not observe the 17 deaths that occurred in each group because many patients were lost to follow-up, or censored, before the end of the study
- The proportion of censoring $(C = 1)$ was higher among patients with heart disease $(L = 1)$ at the start of the study and among those assigned to wasabi supplementation $(A = 1)$.
- 9 individuals in the wasabi group and 22 individuals in the other group were not lost to follow-up

Selection & censoring

- The investigator observed 4 deaths in the wasabi group and 11 deaths in the other group
- the associational risk ratio $Pr[Y =| \, |A = 1, C = 0]/Pr[Y = 1 \mid A = 0, C = 0] = (4/9)/(11/22) = 0.89$ among the uncensored
- selection bias due to conditioning on the common effect C

Selection & censoring



- U: atherosclerosis, affects both heart disease L and death Y
- no shared common causes of A and Y due to marginal randomization, no need to adjust for confounding
- selection due to conditioning on C, possible to block backdoor path $C \leftarrow L \leftarrow U \rightarrow Y$

Selection & censoring

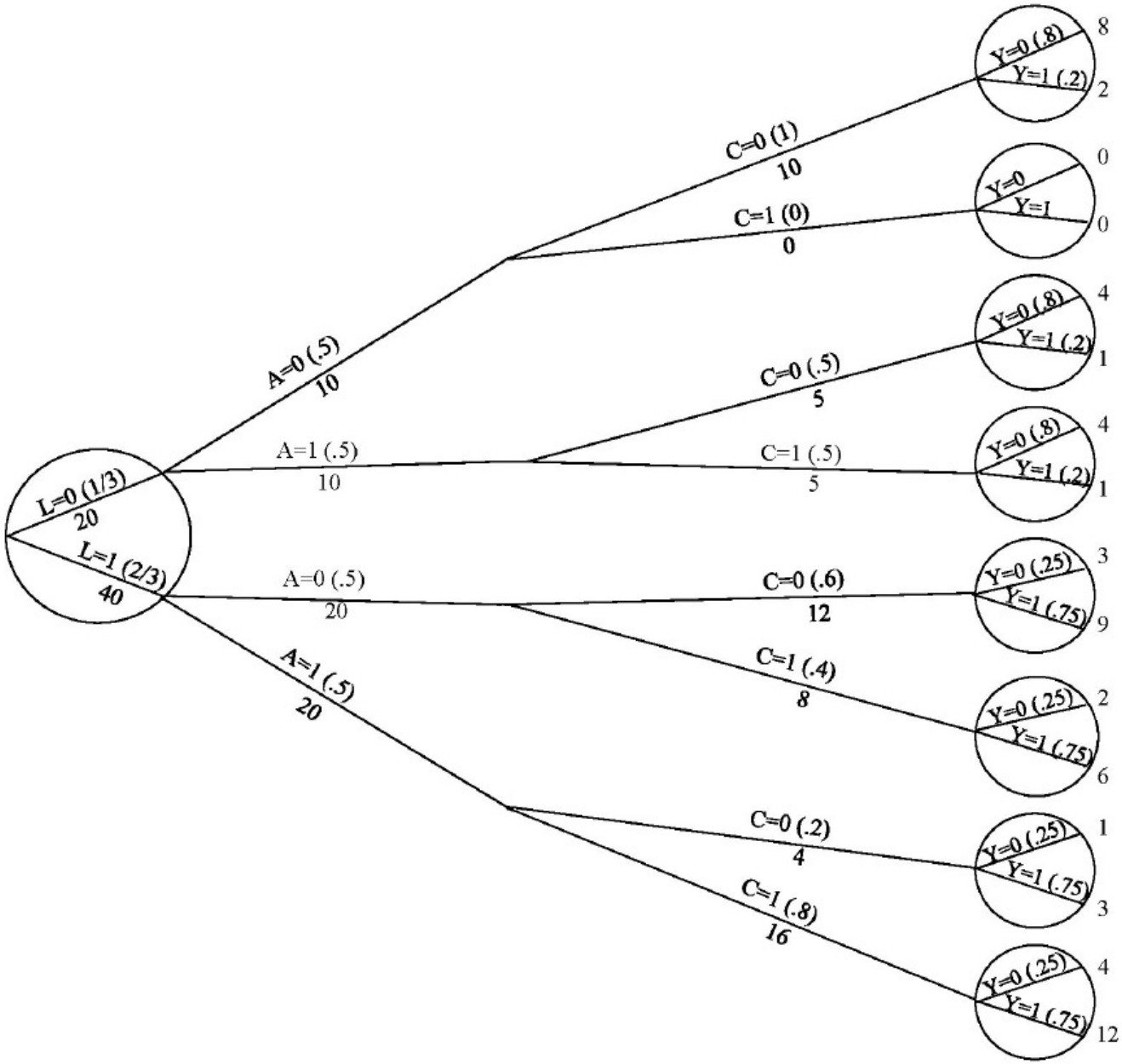
- use $Y^{a=1,c=0}$ denote an individual's counterfactual outcome if he/she had received treatment $A = 1$ and had remained uncensored $C = 0$
- let $Y^{a=0,c=0}$ be an individual's counterfactual outcome if he had not received treatment $A = 0$ and he had remained uncensored $C = 0$.
- causal contrast $Pr(Y^{a=1,c=0} = 1) - Pr(Y^{a=0,c=0} = 1)$
- if censoring does not have effect on outcome, one might ignore C in defining causal effect

Adjust for selection bias

Application of IP weighting to adjust selection bias

- assigning a weight W^C to each selected individual ($C = 0$) so that she accounts in the analysis not only for herself, but also for those like her, i.e., with the same values of L and A, who were not selected ($C = 1$).
- The IP weight W^C is the inverse of the probability of her selection $Pr[C = 0 \mid L, A]$.
- When both confounding and selection bias exist, the product weight $W^A \times W^C$ can be used to adjust simultaneously for both biases under suitable assumptions.

Adjust for selection bias



Adjust for selection bias

20 individuals with heart disease ($L = 1$) who were assigned to wasabi supplementation ($A = 1$)

- 4 remained uncensored, 16 were censored, $Pr(C = 0 \mid L = 1, A = 1) = 0.2$
- in IP weighting, 16 censored individuals receive 0 weight, whereas the 4 uncensored receive a weight of 5 to replace 20 original individuals
- the idea is to construct a pseudo-population of the same size as the original study population but in which nobody is lost to follow-up.

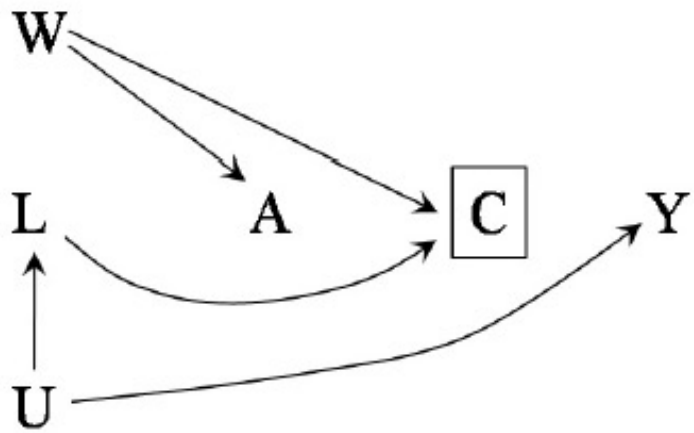
Adjust for selection bias

The association measure in the pseudo-population equals the effect measure in the original population if

- the average outcome in the uncensored individuals must equal the unobserved average outcome in the censored individuals with the same values of A and L. (this will be satisfied if the probability of selection $Pr(C = 0 \mid L = 1, A = 1)$ is conditional on A and all additional factors that independently predict both selection and the outcome)
- IP weighting requires that all conditional probabilities of being uncensored given the variables in L must be greater than zero. Positivity is not required for $C = 1$ since we are not interested in those who censored.
- Intervention is well defined.

Adjust for selection bias

In certain situations, stratification could also work to adjust for selection bias

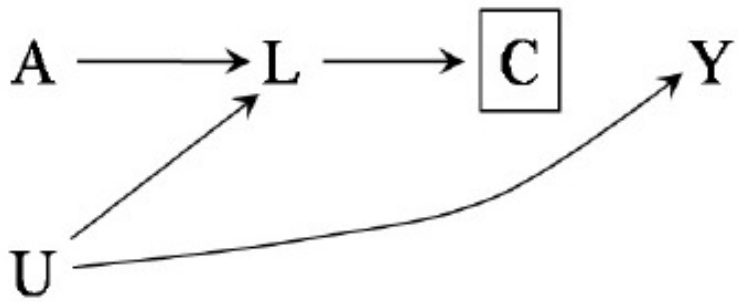


- Stratification could yield unbiased conditional effect measures within levels of L if conditioning on L is sufficient to block the backdoor path from C to Y

$$Pr(Y = 1 \mid A = 1, C = 0, L = l)/Pr(Y = 1 \mid A = 0, C = 0, L = l)$$

Adjust for selection bias

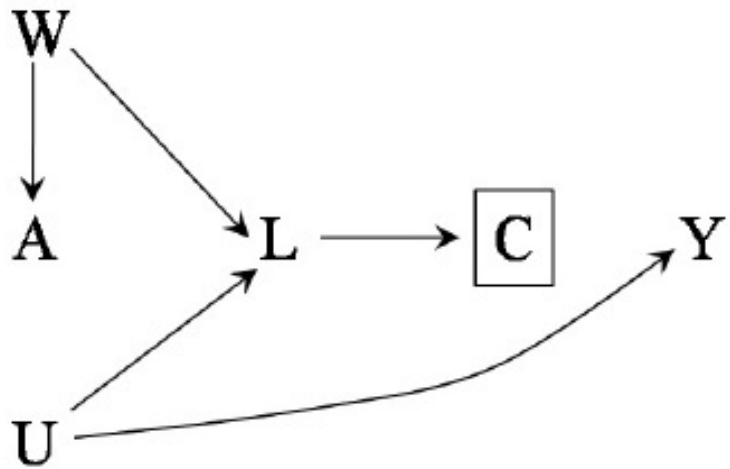
In certain situations, stratification would not work to adjust for selection bias



- because conditioning on L blocks the backdoor path from C to Y
- but opens the path $A \rightarrow L \leftarrow U \rightarrow Y$ from A to Y since L is a collider

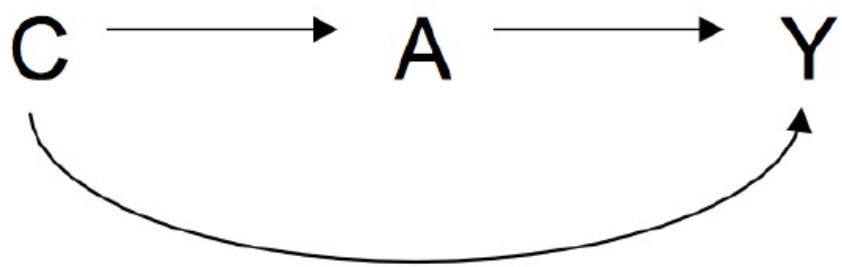
Adjust for selection bias

In certain situations, stratification would not work to adjust for selection bias



- because conditioning on L blocks the backdoor path from C to Y
- but opens the path $A \leftarrow W \rightarrow L \leftarrow U \rightarrow Y$ from A to Y since L is a collider

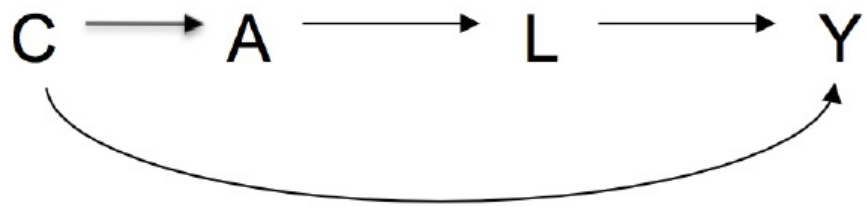
Causal Inference Principle I



- Suppose we wish to estimate the total effect of A on Y
- **Causal Inference Principle I**: If C is a common cause of A and Y, then we should control for C

- If we do not control for C, then the association we observe between A and Y may not be due to the causal effect of A on Y but rather due to the association between A and Y induced by C

Causal Inference Principle II



- Suppose we wish to estimate the total effect of A on Y
- **Causal Inference Principle II**: If there is an intermediate variable L between A and Y, we should not control for it.

- If we do control for L, then some of the association between A and Y due to the causal effect of A and Y may be blocked by controlling for L.

Summary

- Association is not causation and investigators need to be aware under which assumptions statistical analyses yield causal interpretation
- Potential outcomes framework and DAGs help formalizing definition of causal effects, clarifying assumptions, and reason on whether such assumptions are met
- Adjustment for confounding and correct model specification is key to ensure causal interpretation
- Standard regression approaches are typically used for explanatory modeling in the context of fixed time exposures