INVERSE PROBABILITY WEIGHTING
Selection Bias

Barbra Dickerman, Joy Shi, Miguel Hernán
DEPARTMENTS OF EPIDEMIOLOGY AND BIOSTATISTICS
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Learning objectives
At the end of this lecture you will be able to

- Review the conditions for selection bias under the null
- Use IP weighting to adjust for selection bias when estimating causal effects

- Key concepts
  - Selection bias under the null
  - IP weights for selection
  - Differential loss to follow-up
  - Competing risks

Plan for today

A. Review of selection bias

B. IP weighting to adjust for selection bias

- Due to loss to follow-up/missing data

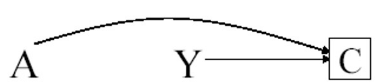
Selection bias

- Ubiquitous concept
- Bias that arises when the parameter of interest in a population differs from the parameter in the subset of individuals from the population that are available for analysis
  - Selection bias for descriptive measures (e.g., prevalence) because of non-random sampling
  - Selection bias for effect measures (e.g., causal risk ratio) because of differential loss to follow-up

Selection bias for effect measures

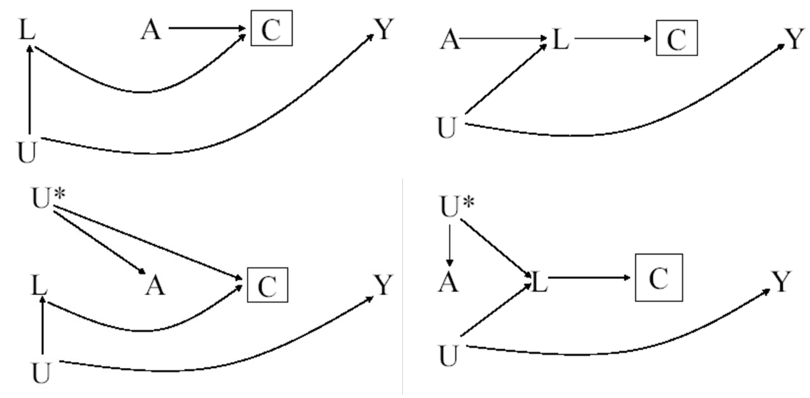
- Many different names
  - Inappropriate selection of controls, Berkson's bias, incidence-prevalence bias, loss to follow-up, nonresponse bias, missing data bias, volunteer bias, self selection, healthy worker effect…
- Here we focus on selection bias that arises even in the absence of a causal effect of treatment on the outcome
  - Selection bias under the null

The structure of selection bias under the null



- The bias arises as the consequence of conditioning on a common effect of treatment and outcome
  - or on a common effect of a cause of the treatment and a cause of the outcome
- That is, the design or the analysis is conditioned on "being selected for analysis" C=0

Missing data / Nonresponse
Censoring / Loss to follow-up



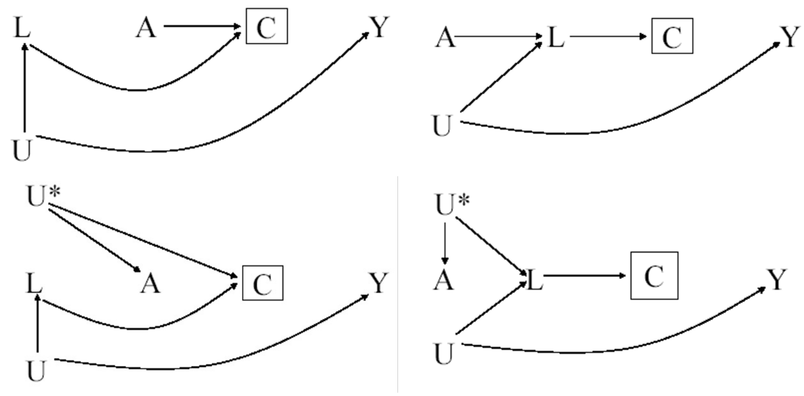
C: Missing data (1: yes, 0: no)

Bias due to differential loss to follow-up is possible in randomized experiments

True 0%

False 0%

Volunteer bias / Self-selection bias



Bias due to self-selection of participants at baseline is possible in randomized experiments

0

True

0%

False

0%

Aside: Internal vs. external validity in randomized experiments

- Internal validity
  - the estimated association has a causal interpretation in the studied population
  - i.e., no selection bias, no confounding
- External validity
  - the estimated association has a causal interpretation in another population
  - i.e., generalizability or transportability

- In randomized experiments
  - There is internal validity
  - Perhaps not external validity

A note on terminology

- The usage of the terms ‘confounding’ and ‘selection bias’ is not standardized

- Epidemiologists use ‘confounding’ and statisticians/econometricians ‘selection bias’ when referring to the same bias
- Others use ‘selection bias’ when ‘confounders’ are unmeasured
- Some use the term ‘selection-confounding’

A note on terminology

- We refer to the presence of common causes as confounding, and to conditioning on common effects as selection bias
  - This classification may not coincide perfectly with the traditional, often discipline-specific, terminologies
- Our goal is not to be normative about terminology
  - but rather to emphasize that there exist two distinct causal structures that lead to bias
  - regardless of the terms chosen to refer to them

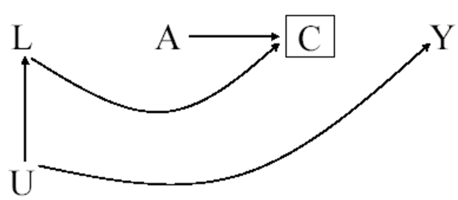
Selection bias in our study population?

- 1629 cigarette smokers
- Aged 25-74 years when interviewed in 1971-75 (baseline)
- Interviewed again in 1982
- Known sex, age, race, weight, height, education, alcohol use, and smoking intensity at both baseline and follow-up visits, and who answered the general medical history questionnaire at baseline

Key variables

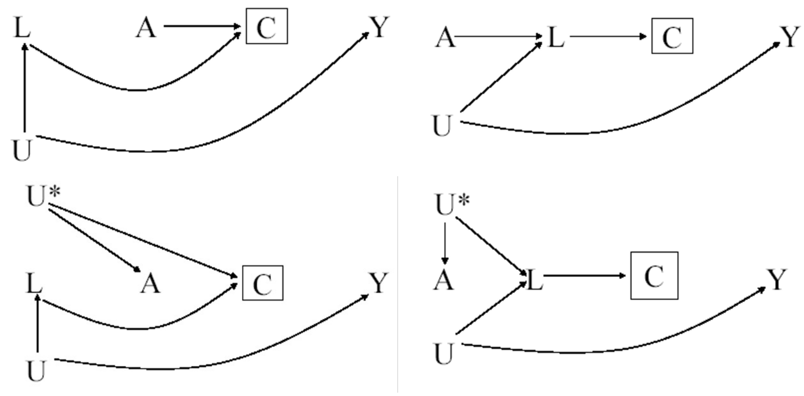
Treatment A	Quit smoking between baseline and 1982 1: yes, 0: no
Continuous outcome Y	Weight gain, kg Weight in 1982 minus baseline weight Available for 1566 individuals
Dichotomous outcome D	Death by 1992 1: yes, 0: no
Baseline (pre-treatment) covariates	Age, sex, race, alcohol use, intensity of smoking, weight…
Censoring C	Missing weight in 1982 1: yes, 0: no

Differential loss to follow-up / nonresponse or missing data



A: Smoking cessation, Y: weight gain,
C: Censoring (1: yes, 0: no),
L: Smoking intensity in 1971-75,
U: Lifetime history of smoking

Differential loss to follow-up



C: Missing data (1: yes, 0: no)

Plan for today

A. Review of selection bias

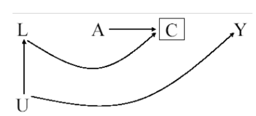
B. IP weighting to adjust for selection bias

- Due to loss to follow-up/missing data

Adjustment for selection bias

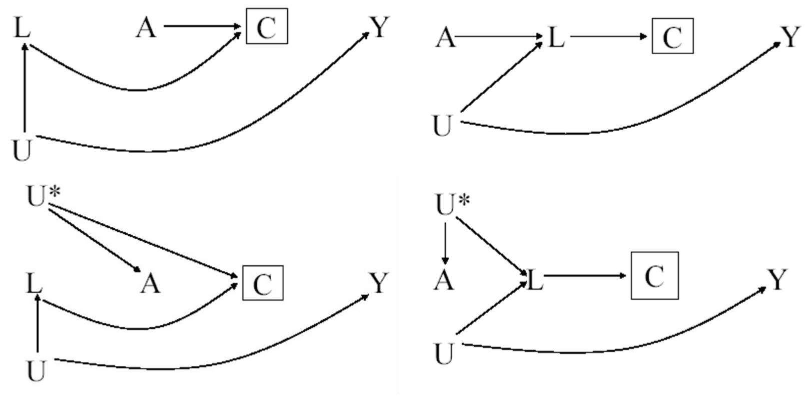
- Sometimes selection bias can be prevented by study design
  - e.g., sampling controls in a manner to ensure that they will represent the treatment distribution in the population
- Most often selection bias needs to be adjusted for via
  - Stratification
  - IP weighting

Stratification to adjust for selection bias



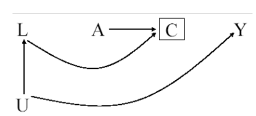
- The idea is blocking the path that was unblocked because of conditioning on the collider C
  - Can block it if L (or U) is measured
- The path is blocked by estimating the A-Y association in the selected within levels of L
  - That is, adding a box around L
- The conditional association measure is the causal effect within levels of L and C=0
  - a bit weird but unbiased under the null

Stratification to adjust for selection bias?



Conditional association measure not always causal!

IP weighting to adjust for selection bias

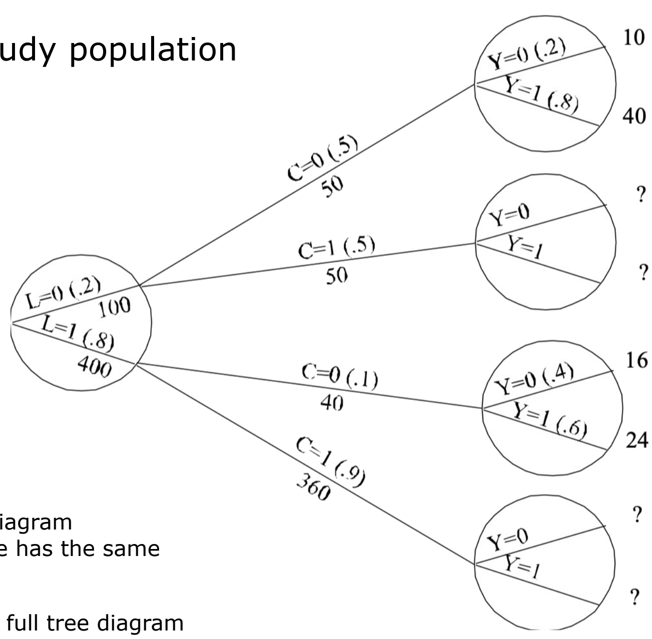


- ☐ The idea is eliminating the path that was unblocked because of conditioning on the collider C
  - ■ Can eliminate it if L (or U) is measured
- ☐ The path is eliminated by creating a pseudo-population in which everybody is selected (e.g., uncensored)
  - ■ Because then there are no arrows into C
- ☐ The association measure in the pseudo-population is the effect measure in the study population
  - ■ Unconditionally

Example

- 100 participants with same treatment and covariate history
  - untreated, men, aged 40-45, CD4 count >500
- 50 are lost to follow-up and do not contribute to the analysis (zero weight)
- The remaining 50 receive a weight=2
  - Probability of remaining uncensored is 0.5, weight for uncensored individuals is 1/0.5=2
  - IP weighting creates a pseudo-population in which the 100 participants are replaced by 2 copies of the 50 uncensored individuals

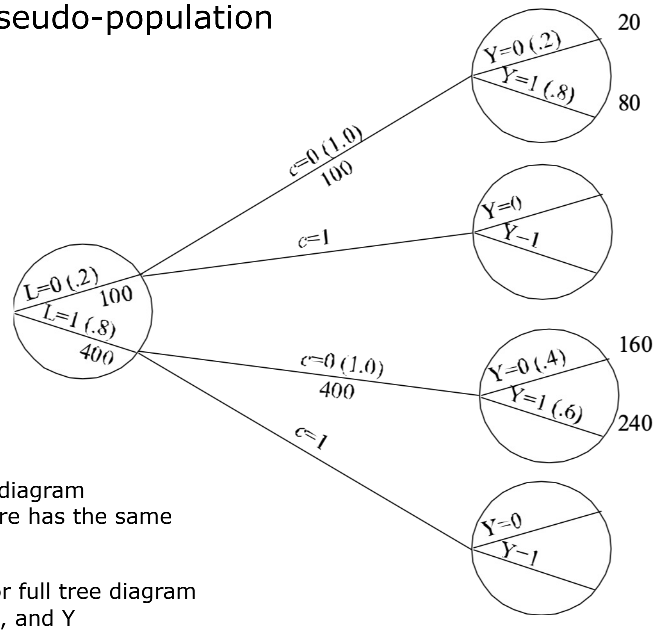
Study population



Note: Simplified tree diagram (assume everyone here has the same treatment A)

See course website for full tree diagram with nodes for L, A, C, and Y

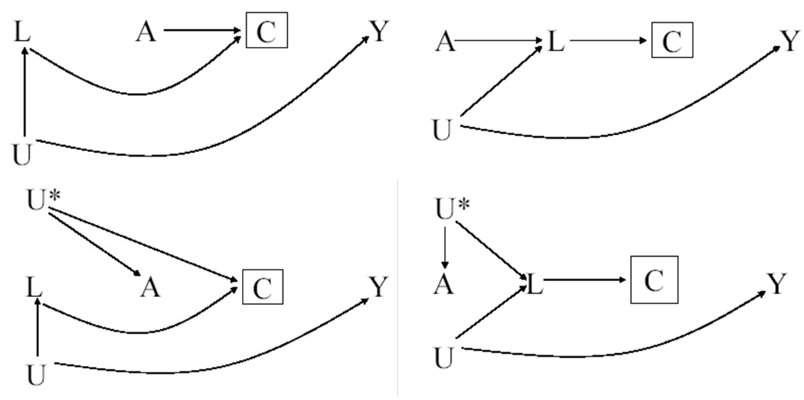
Pseudo-population



Note: Simplified tree diagram
(assume everyone here has the same treatment A)

See course website for full tree diagram with nodes for L, A, C, and Y

IP weighting to adjust for selection bias



Works in all four cases by removing arrows into *C*

Which assumption are we making?

$Y^{a,c=0} \amalg C|A,L$ for $c=0$

- Conditional exchangeability
  - within levels of A,L the risk in the unselected if selected is the same as the risk in the selected
  - or selection is randomized within levels of A,L
  - or no unmeasured confounding for selection C within levels of the measured variables A,L
- Also required:
  - Positivity: not all censored for some A,L levels
  - Consistency, including a well-defined intervention to eliminate censoring
    - problems with competing risks (see later)

IP weights for selection bias

$$W^C = \frac{1}{\Pr[C = 0|A, L]}$$

- Each selected individual in the population is weighted to create W^C individuals in the pseudo-population
  - Unselected individuals have weight zero
- The denominator of your weight is (informally) the probability of having been selected given your A,L values
  - Equal for all $C=0$ individuals with same A,L values

IP weights

- To adjust for confounding
  - Use IP weights W^A
  - A is the treatment variable
- To adjust for selection bias
  - Use IP weights W^C
  - C is the selection variable
- To adjust for both biases
  - Multiply $W^A \times W^C$

Stabilized IP weights for selection bias

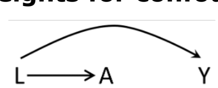
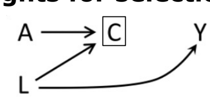
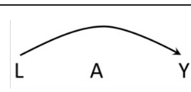
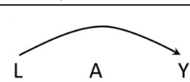
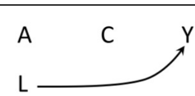
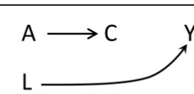
$$SW^C = \frac{\Pr[C = 0|A]}{\Pr[C = 0|A, L]}$$

- Same denominator as nonstabilized IP weights multiplied by an individual's probability of having been selected given their *A* values
- Each selected individual *i* in the population is weighted to create SW_i^C individuals in the pseudo-population

Stabilized IP weights

- To adjust for confounding
  - Use IP weights SW^A
  - A is the treatment variable
- To adjust for selection bias
  - Use IP weights SW^C
  - C is the selection variable
- To adjust for both biases
  - Multiply $SW^A \times SW^C$

Summary of IP weights for confounding and selection bias

	IP weights for confounding 		IP weights for selection bias 	
	Nonstabilized	**Stabilized**	**Nonstabilized**	**Stabilized**
Formula	$\frac{1}{f(A\|L)}$	$\frac{f(A)}{f(A\|L)}$	$\frac{1}{\Pr[C=0\|A,L]}$	$\frac{\Pr[C=0\|A]}{\Pr[C=0\|A,L]}$
DAG for pseudopopulation				
Size of pseudopopulation	Original N * number of levels of A	Original N	Original N before censoring	Original N after censoring

Nonstabilized weights eliminate censoring

0

True

0%

False

0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Stabilized weights eliminate censoring

True

0%

False

0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Back to our data
Causal question 1

- Estimate the mean weight gain if everybody had quit smoking
  - $E[Y^{a=1}]$
- Estimate the mean weight gain if nobody had quit smoking
  - $E[Y^{a=0}]$
- Estimate the average causal effect on the additive scale
  - $E[Y^{a=1}] - E[Y^{a=0}]$

We have ignored censoring all this time!

- We restricted the analysis to the 1566 participants with non missing outcome
  - i.e., we did not use data from the 63 participants with missing outcome
- Effectively we assumed that the 1566 uncensored and the 63 censored participants were exchangeable
  - We assumed no selection bias due to censoring C

See ipw_selection.R, lines 6-9

Causal question 1 (what we really meant)

- Estimate the mean weight gain if everybody had quit smoking **and nobody had been censored**
  - $E[Y^{a=1,c=0}]$
- Estimate the mean weight gain if nobody had quit smoking **and nobody had been censored**
  - $E[Y^{a=0,c=0}]$
- Estimate the average causal effect
  - $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$

What model were we really fitting?

- We fit the weighted regression model

  $E[Y|A]= \theta_0 + \theta_1 A$
  - but restricted to individuals with nonmissing weight gain, i.e., conditional on C=0
- That is, we were really fitting the weighted regression model

  $E[Y|A,C=0]= \theta_0 + \theta_1 A$

What if this were a randomized experiment…

- … in which individuals had been randomly assigned to "quit smoking" or "remain as smokers"?
- Is the difference $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$ consistently estimated by $\hat{\theta}_1$ from the unweighted regression model?

  $\mathrm{E}[Y|A,C=0]= \theta_0 + \theta_1 A$
- Not in general
  - Post-randomization loss to follow-up/missing data may destroy the baseline exchangeability achieved by randomization

What if this were an observational study…

- … in which the treated and the untreated are exchangeable **at baseline**, conditional on measured confounders?
  - e.g., age, sex, race, alcohol, etc.
- IP weighting creates a pseudo-population with unconditional exchangeability **at baseline**
- In general, $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$ is not consistently estimated by $\hat{\theta}_1$ from the weighted regression model

  $\mathrm{E}[Y|A,C=0]= \theta_0 + \theta_1 A$

Which assumption were we making?

- Exchangeability
  - unconditional with respect to covariates L
  - selection is randomized within levels of A only
- This assumption is stronger than the assumption of conditional exchangeability
  - selection is randomized within levels of A,L
- Let's then conduct the analysis under the weaker assumption by using IP weighting

Doubly-weighted regression model

$E[Y|A,C=0]= \theta_0 + \theta_1 A$

- Adjusting for the following variables
  - Sex, age, race, hbp, education, active, smokeyrs, smokeintensity, exercise
    - squared terms for continuous variables
  - For both SW^A and SW^C
- Parameter estimates
  - $\hat{\theta}_0$ = 1.8
  - $\hat{\theta}_1$ = 3.3 (conservative 95% CI: 2.3, 4.3)
- Saturated model
  - 2 parameters, 2 quantities

See ipw_selection.R, lines 11-64

Censoring and the g-formula

- Did we adjust for censoring when using standardization?
- See homework

Summary: IP weighting adjusts for selection bias

- By creating a pseudo-population in which
  1. selection (e.g., censoring, missing data) is eliminated or randomly allocated
  2. the effect of the treatment is the same as in the original population
- The pseudo-population effect measure is equal to the effect measure had everybody been selected in the original population

Interpretation for different types of selection

- Censoring due to loss to follow-up
  - Effect had nobody, or only a random sample, been lost to follow-up
  - Appropriate
- Missing data, nonresponse
  - Effect had nobody, or only a random sample, had missing data
  - Appropriate
- Censoring due to competing risks
  - Effect had nobody, or only a random sample, been censored due to competing risks
  - Probably not interesting

Example of competing risks

- A study to estimate the effect of cigarette smoking on the risk of Alzheimer's disease

- Do we want effect estimates from a pseudo-population in which all other causes of death (cancer, heart disease, stroke, etc.) have been removed?
  - Pseudo-population does not correspond to any known human population
  - Plus, no well-defined intervention could possibly remove just one cause of death without affecting the others as well

Readings

- *Causal Inference, What If*. Chapter 8

- Greenland S. Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology* 2003; 14:300-306
- Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; 15:615–625

Progress report

1. Introduction to modeling
2. Stratified analysis:
   - outcome regression
   - propensity scores
3. Standardization
4. Inverse probability weighting
   - Marginal structural models
5. Instrumental variable estimation