University of California
San Francisco

Selection Bias

Epi 265

May 9, 2016

Elizabeth Rose Mayeda, PhD, MPH
Postdoctoral Fellow
University of California, San Francisco
Department of Epidemiology and Biostatistics

Quiz

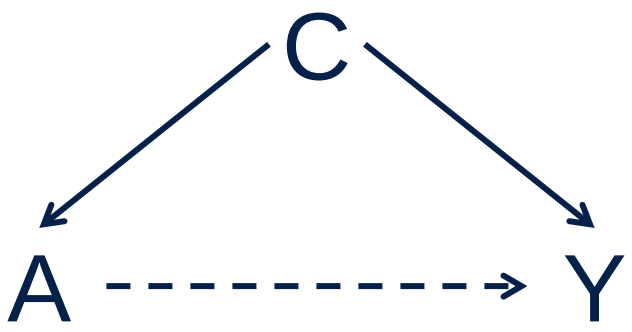
- Consider a hypothetical trial of a new drug treatment (A) to reduce the probability of memory impairment (Y) assessed 1 year after randomization to the drug. Write the ITT estimate for the risk difference, assuming perfect adherence, everyone survives and continues to participate through the 1 year outcome assessment:
  - Write the counterfactual contrast estimated by the above ITT.
  - Can you infer that the ITT estimates the Population Average Treatment Effect and do you need the assumptions of exchangeability, consistency, or positivity to draw this inference?
- Now imagine that drug A reduces survival to 1 year follow-up: $Pr[S^{a=1}=1]<Pr[S^{a=0}=1]$. Write out the best ITT estimate of the risk difference, assuming you can only evaluate cognitive impairment on people who remain alive.
  - Write the counterfactual contrast estimated by the above.
  - Can you infer that the ITT estimates the Population Average Treatment Effect and do you need the assumptions of exchangeability, consistency, or positivity to draw this inference?

Outline

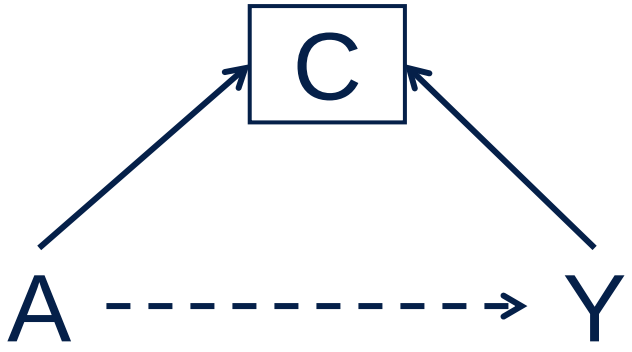
- Introduction to selection bias
  - Causal structures of selection bias
- Examples of selection bias in:
  - Case-control studies
  - Cohort studies
  - RCTs
- Tools to address selection bias (focus on IPW)
- Quantifying survival bias in research on determinants of cognitive decline with a simulation framework

What is selection bias?

- **Confounding:** Bias arising from the existence of common causes of exposure and outcome



- **Selection bias:** Bias arising from conditioning on a common effect (collider) of exposure (or cause of exposure) and outcome (or cause of outcome)



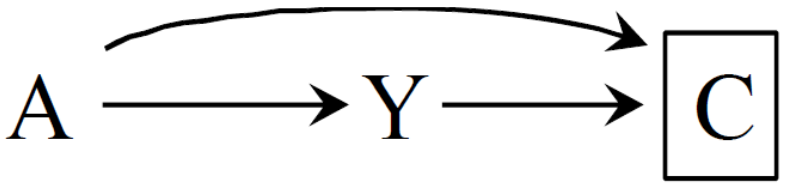
What is selection bias?

- Arises from differential:
  - Selection into study (by design or self-selection by participants)
  - Selective attrition (drop out/death) of enrolled participants
  - Nonresponse, missing data
- Consequence: Association between exposure and outcome among the selected study participants differs from those eligible
- Both observational and randomized studies are vulnerable to selection bias
- Causal diagrams are helpful for depicting causal structures that give rise to selection bias

Causal structures of selection bias

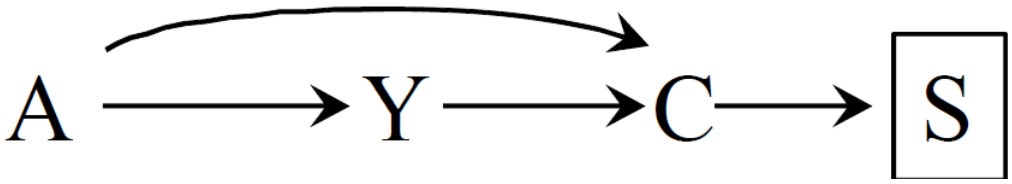
- Conditioning on a common effect of exposure and outcome results in two sources of association between exposure and outcome:
  1. Causal effect of A on Y: A→Y
  2. Spurious association between A and Y induced by conditioning on collider C: A→C ← Y

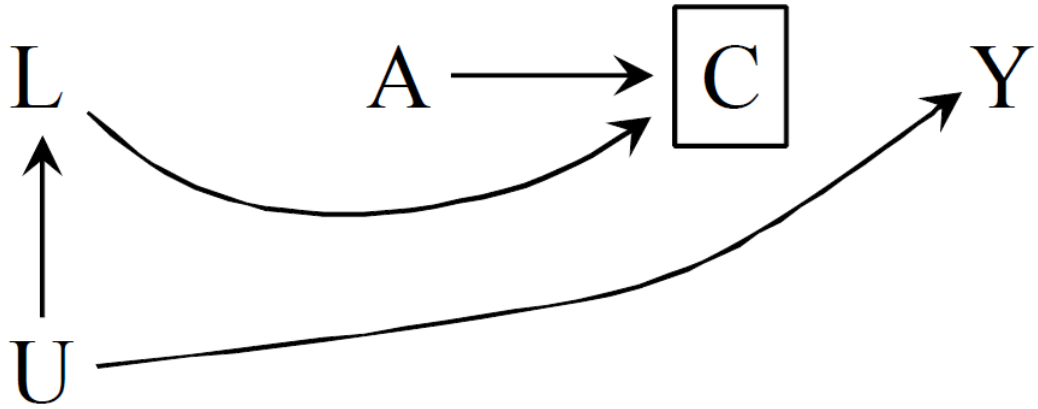
- Because of selection bias, association ≠ causation:
  - $Pr[Y=1|A=1,C=0]/Pr[Y=1|A=0,C=0] \neq Pr[Y^{a=1}=1]/Pr[Y^{a=0}=1]$



Causal structures of selection bias

- Extension: Conditioning on a common effect of exposure (or cause of exposure) and outcome (or cause of outcome) induces selection bias

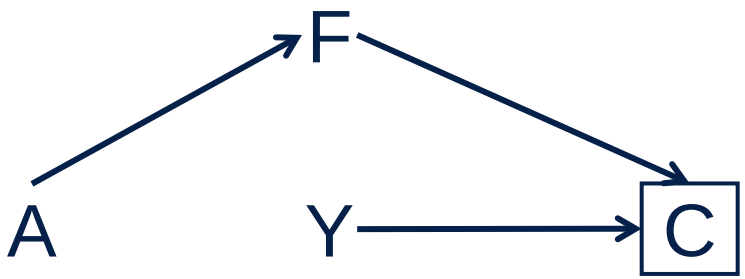




Selection bias in case-control studies

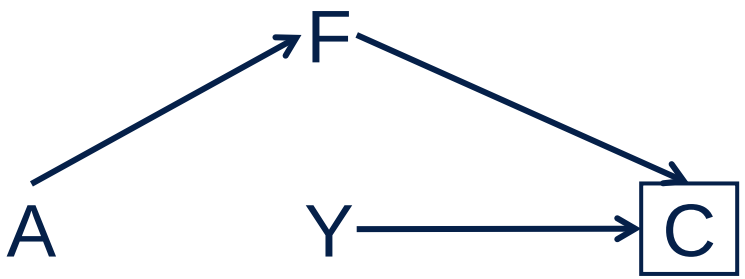
Case-control studies: inappropriate selection of controls

- Case-control study of the effect of HRT on myocardial infarction in postmenopausal women
  - A: Hormone replacement therapy (HRT)
  - Y: Coronary heart disease
  - F: Hip fracture
  - C: Selection into study



Case-control studies: incidence-prevalence bias

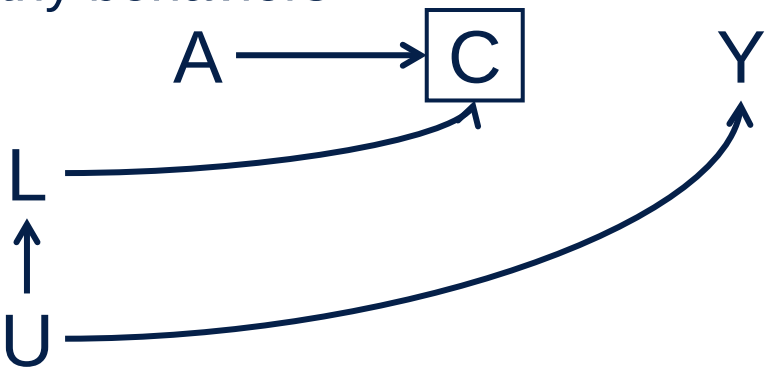
- Case-control study of cigarette smoking and Parkinson's disease in middle-aged and older adults
  - A: Cigarette smoking
  - Y: Parkinson's disease
  - F: CVD, cancer
  - C: Study restricted to people alive at a certain age



Selection bias in cohort studies and RCTs

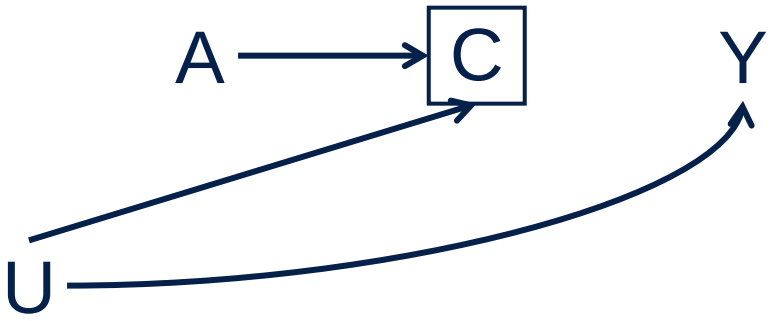
Selection bias in cohort studies: selection into cohort

- Cohort study of cigarette smoking and coronary heart disease among middle-aged adults
  - A: Cigarette smoking
  - Y: Coronary heart disease
  - C: Volunteered to participate
  - U: Family history of coronary heart disease
  - L: Heart healthy behaviors



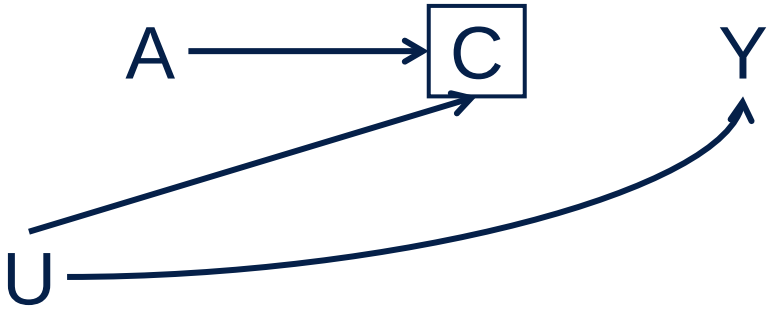
Selection bias in cohort studies: selection into cohort

- Cohort study of cigarette smoking and cognitive decline among the oldest-old (age 85+)
  - A: Cigarette smoking
  - Y: Cognitive decline
  - C: Alive and healthy enough to participate in study at age 85+
  - U: Unmeasured common causes of survival to old age and cognitive decline



Selection bias in cohort studies & RCTs: selective attrition

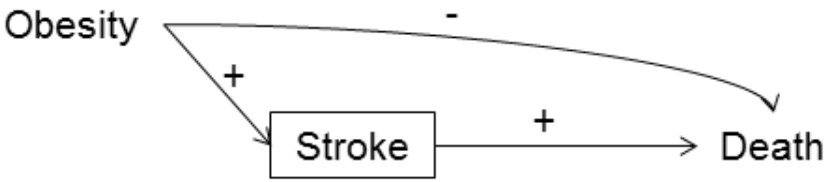
- Selection occurs after randomization, so randomization does not protect against this source of selection bias
- RCT of intensive glucose control treatment among adults with type 2 diabetes
  - A: Intensive glucose control treatment
  - Y: Cognitive decline
  - C: Survival
  - U: Unmeasured common causes of survival and cognitive decline



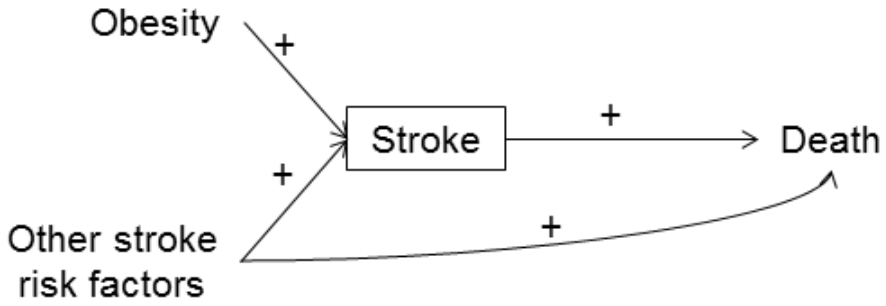
Selection bias arising from conditioning on a disease in a cohort study

- Obesity predicts excess morbidity and mortality in the general population, but among many chronic disease patients (including, stroke, heart failure, and cancer), obesity are associated with *lower* risk of mortality
  - Should obese patients with these chronic conditions be counseled to lose weight?
  - Should normal weight patients with these chronic conditions be counseled to gain weight?

Causal scenario



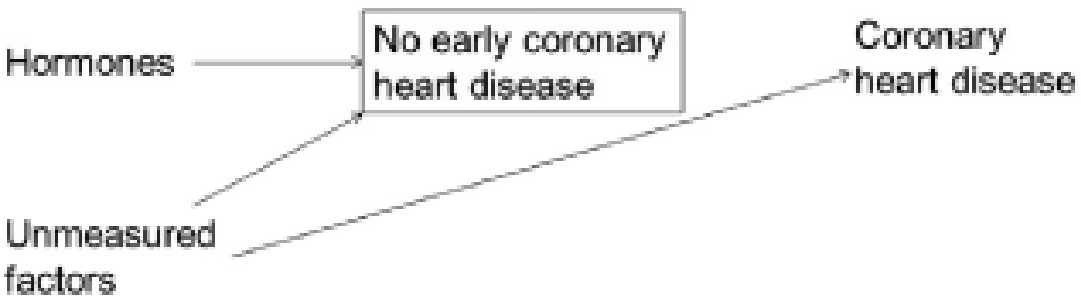
Selection bias scenario



UCSF

Selection bias arising from gap between time of start of exposure and follow up

- Discrepancy between evidence on effect of HRT on CHD from observational studies and RCTs:
  - Observational studies (HRT protective): Selected prevalent HRT users without prevalent CHD and followed them for incident CHD
  - RCTs (HRT harmful): Randomized women without prevalent CHD to HRT or placebo and followed them for incident CHD
- Observational studies effectively ignored the first years of follow up by comparing prevalent users to never users

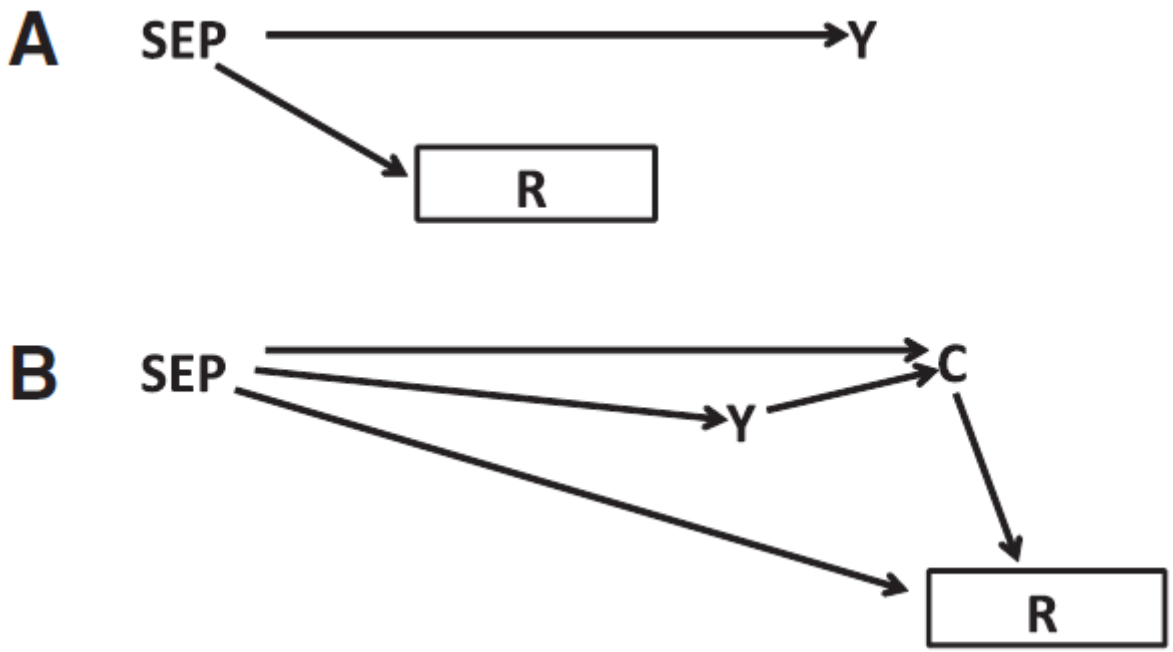


Example: Socioeconomic health inequalities

- Cohort members tend to be more affluent and healthier compared to the whole population, but this does not bias certain exposure-outcome associations
- Howe *et al*: How does selected participation influence estimates of effects of socioeconomic health inequalities?
- Avon Longitudinal Study of Parents and Children
  - Estimated effect of maternal education on outcomes in a) the whole cohort (n≈12,000) and b) restricted to subsamples participating at age 10 (n ≈ 7,000) and age 15 (n ≈ 5,000)
- Assuming SEP influences participation, what are some causal structures where we would and would not observe biased estimates of effects of SEP on health?

Example: Socioeconomic health inequalities

- SEP=exposure of interest, Y=outcome of interest, R=response, C=conditions that influence participation (e.g. illness making participation more difficult or health awareness increasing participation)



Example: Socioeconomic health inequalities

- Continued participation associated with both SEP and outcomes
- Loss to follow-up resulted in *underestimation* of inequalities, although qualitative conclusions held for most outcomes. For example, mean birth weight difference between highest and lowest SEP groups was:
  - Full sample: 116 grams (95% CI: 78-153)
  - Sample participating age 10: 93 grams (95% CI: 45-141)
  - Sample participating age 15: 62 grams (95% CI: 5-119)

Tools to address selection bias

Tools to address selection bias

Repeated measures

- Study design
- Adjustment
- IPW
- Joint longitudinal-survival models
- Multiple imputation
- Sensitivity analysis and bias correction (estimate what the bias would be from your model and then subtract it)

Time to event

- Study design
- Adjustment
- IPW
- Fine-gray
- Multistate models
- Multiple imputation
- Sensitivity analysis and bias correction (estimate what the bias would be from your model and then subtract it)

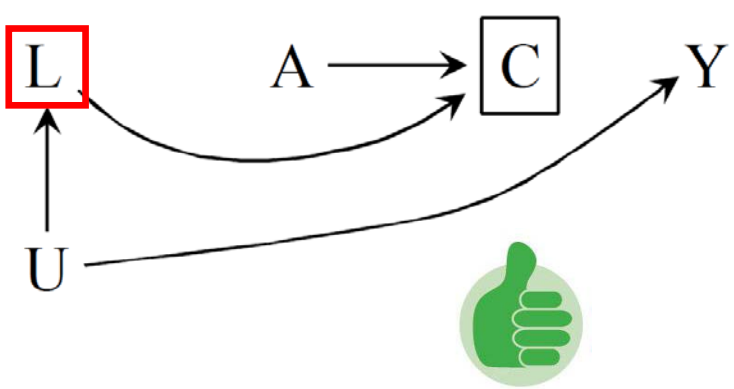
Study design to address selection bias

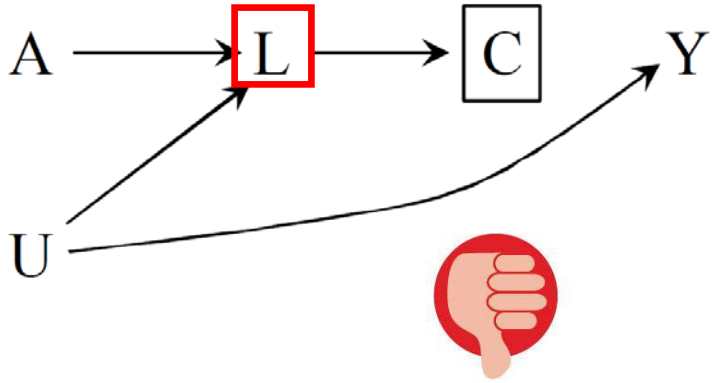
- Selection of appropriate controls in a case-control study
- Avoid gap between time of start of exposure and follow up (incident user design) in a longitudinal study
- Measures to reduce missing data and attrition between exposure and outcome assessment in a longitudinal study

Adjustment to address selection bias

- Depending on the causal structure, stratification/adjustment can *sometimes* be used to control for selection bias
- Idea is to block the path that was unblocked by conditioning on the collider (C)

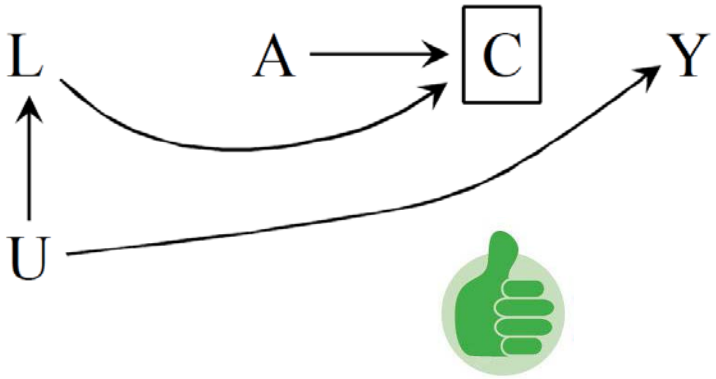
Can fit model $E[Y| Z,L,C=0]=b_0+b_1A+b_2L$ to estimate conditional (on L and C=0) association between A and Y

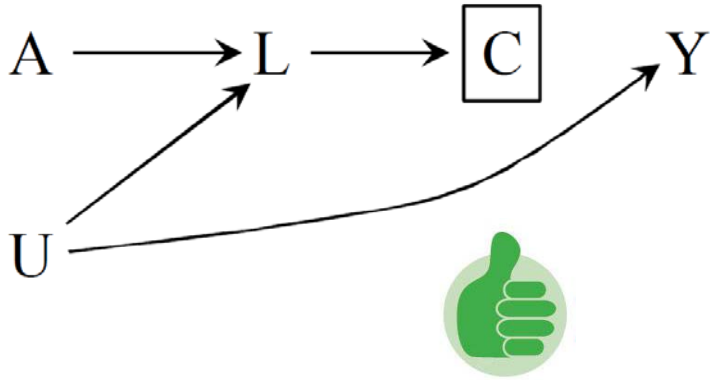




Inverse probability weighting (IPW) to address selection bias

- Goal is to block the path that was unblocked by conditioning on the collider (C) (block path that confounds association between C and Y)
- "Treatment" for IPW to address selection bias is C (unlike when you use IPW to address confounding, where A is the treatment)





IPW to address selection bias

- We assign a weight, W, to each selected/noncensored (C=0) individual so that she represents not only herself but also individuals like her (i.e., same values of A and L) who were not selected (C=1); the weights create a pseudo-population
- Assumes exchangeability of censored and uncensored, conditional on A, L ($Y^c \amalg C|A=a,L=l$)
- A selected individual's weight is the inverse of the probability of her selection given her values of A,L (P[C=0|A,L])
- Nonselected (C=1) individuals have weight zero
- For stabilized weights, numerator is the individual's probability of having been selected given her A values
- Nonstabilized weights: $W = \frac{1}{P[C=0,A,L]}$
- Stabilized weights: $SW = \frac{P[C=0,A]}{P[C=0,A,L]}$

Specifying counterfactual contrast of interest guides analysis

- An essential, but often overlooked, step in conducting meaningful research is to clearly define the estimand of interest
- *"No method is so sophisticated that it can relieve the investigator of the obligation to provide a well-formulated causal question in the first place"* -Kaufman JS, Hernán MA**.** Commentary: Epidemiologic Methods Are Useless: They Can Only Give You Answers. Epidemiology. 2012;23(6):785-6
- Counterfactual contrast (and appropriate statistical method) depends on the research question
- With respect to selection bias, clearly defining the estimand of interest is particularly relevant for selective survival

Estimating effects in selected survivors: 2 estimands

1. Effect of exposure on proportion of the initial (alive, dementia-free) population who developed dementia within a specified period of follow-up time after exposure (i.e., population burden of dementia estimated using Fine and Gray competing risk regression)
   - Difference in cumulative incidence of dementia over a specified follow-up time after exposure: $Pr(Y_{x=1}=1|S_{t=0}=1) - Pr(Y_{x=0}=1|S_{t=0}=1)$ during a specified time period
   - The numerator is the number of people who developed dementia during the specified period of follow-up time after exposure and the denominator is the population that was alive and free of dementia *at the time of exposure*.

Estimating effects in selected survivors: 2 estimands

2. Effect of exposure on dementia among people who would survive under either exposure status (survivor average causal effect (SACE) estimated using IPW)

   - Risk difference among the survive under either treatment group: $[Pr(Y_{A=1}) – Pr(Y_{A=0})] \mid [S_{A=1}=S_{A=0}=1]$
   - It is this contrast between the SACE and estimands calculated among survivors that leads to the label “survival bias,” in that there is an association between exposure and outcome (among survivors) even when exposure has no impact on an individual’s probability of developing dementia

Survivor average causal effect (SACE)

- The SACE was developed to address the conundrum that for individuals who would die under one exposure status, the effect of exposure on the outcome is undefined
- Drawing on a principal stratification framework, the population is conceptualized as including four types of individuals:

1. Those who would survive up to time t under either exposure status (always survive)
2. Those who would die by time t under either exposure status (never survive)
3. Those would survive up to time t if exposed but die otherwise (survive only if X=1)
4. Those who would die by time t if exposed but survive otherwise (survive only if X=0)

Tchetgen Tchetgen EJ, Phiri K, Shapiro R**.** A Simple Regression-based Approach to Account for Survival Bias in Birth Outcomes Research. Epidemiology. 2015;26(4):473-80.

Sensitivity analysis using episensi macro

- Hypothetical case-control study of occupational exposure to resins and lung-cancer death. Selection bias arises if selection probabilities among cases and non-cases differs by exposure status:

	Exposed	Unexposed	Total
Cases	45	94	139
Controls	257	945	1202
Total	302	1039	1341

```
. episensi 45 94 257 945, st(cc) dpscex(c(.9)) dpscun(c(.5)) dpsnex(c(.5))
> dpsnun(c(.7))

Pr Case Selection Exposed: Constant(.9)
Pr Case Selection No-Exposed: Constant(.5)
Pr No-Case Selection Exposed: Constant(.5)
Pr No-Case Selection No-Exposed: Constant(.7)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
   External adjusted Odds Ratio = 0.70
   Percent bias = 152%
```

Selection bias factor: (0.9/0.5)/(0.5/0.7)=2.5

Selection bias adjusted OR: 1.76/2.5=0.70

Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. Stata Journal. 2008;8(1):29-48

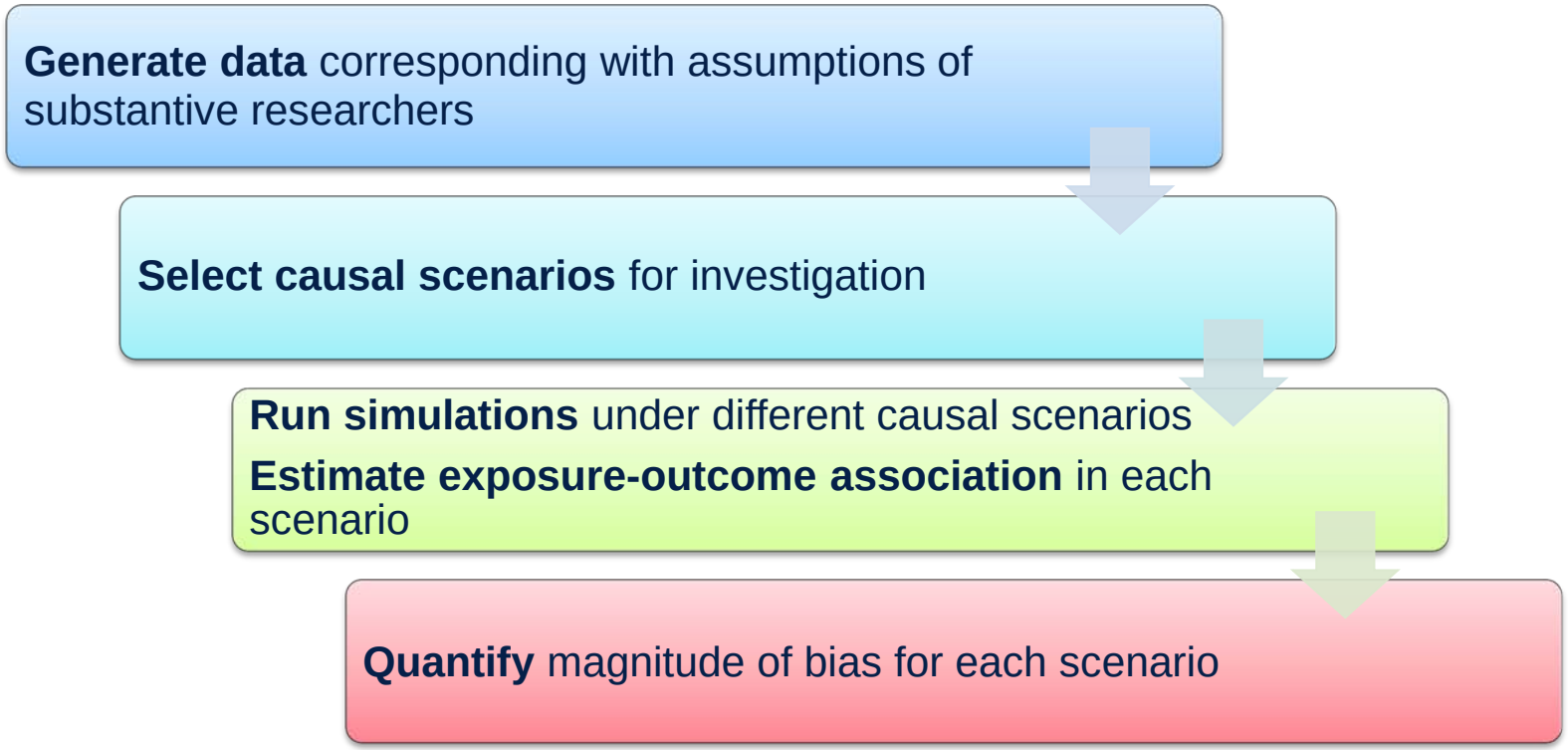
A simulation platform to quantify survival bias: an application to research on determinants of cognitive decline

A simulation platform to quantify survival bias

- Survival bias is a major concern for researchers studying determinants of cognitive decline because many exposures of interest for dementia researchers influence mortality and cognitive decline predicts mortality
- Difficult to anticipate the degree of bias in empirical studies
- Simulation studies provide a means to quantify expected degree of bias in studies of determinants of cognitive decline

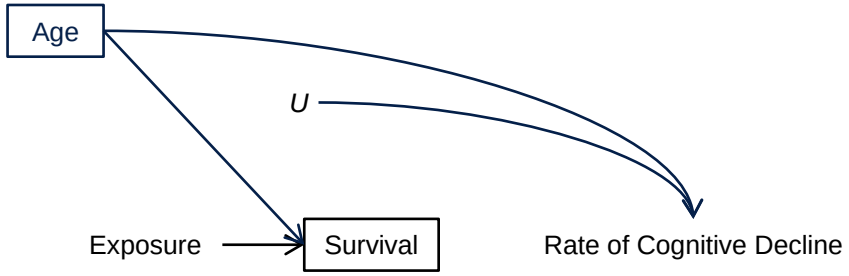
A simulation platform to quantify survival bias

- Simulations provide a means to quantify degree of expected bias in research on determinants of cognitive decline under several causal scenarios



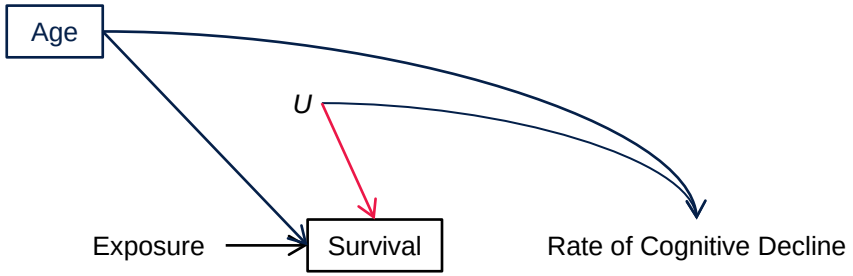
A simulation platform to quantify survival bias

Scenario A (no anticipated bias)

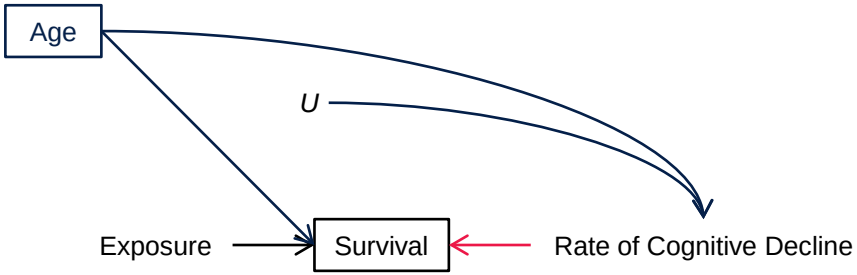


Scenario B1 exposure and U affect mortality

Scenario B2 exposure and U have more than multiplicative effects on mortality hazard

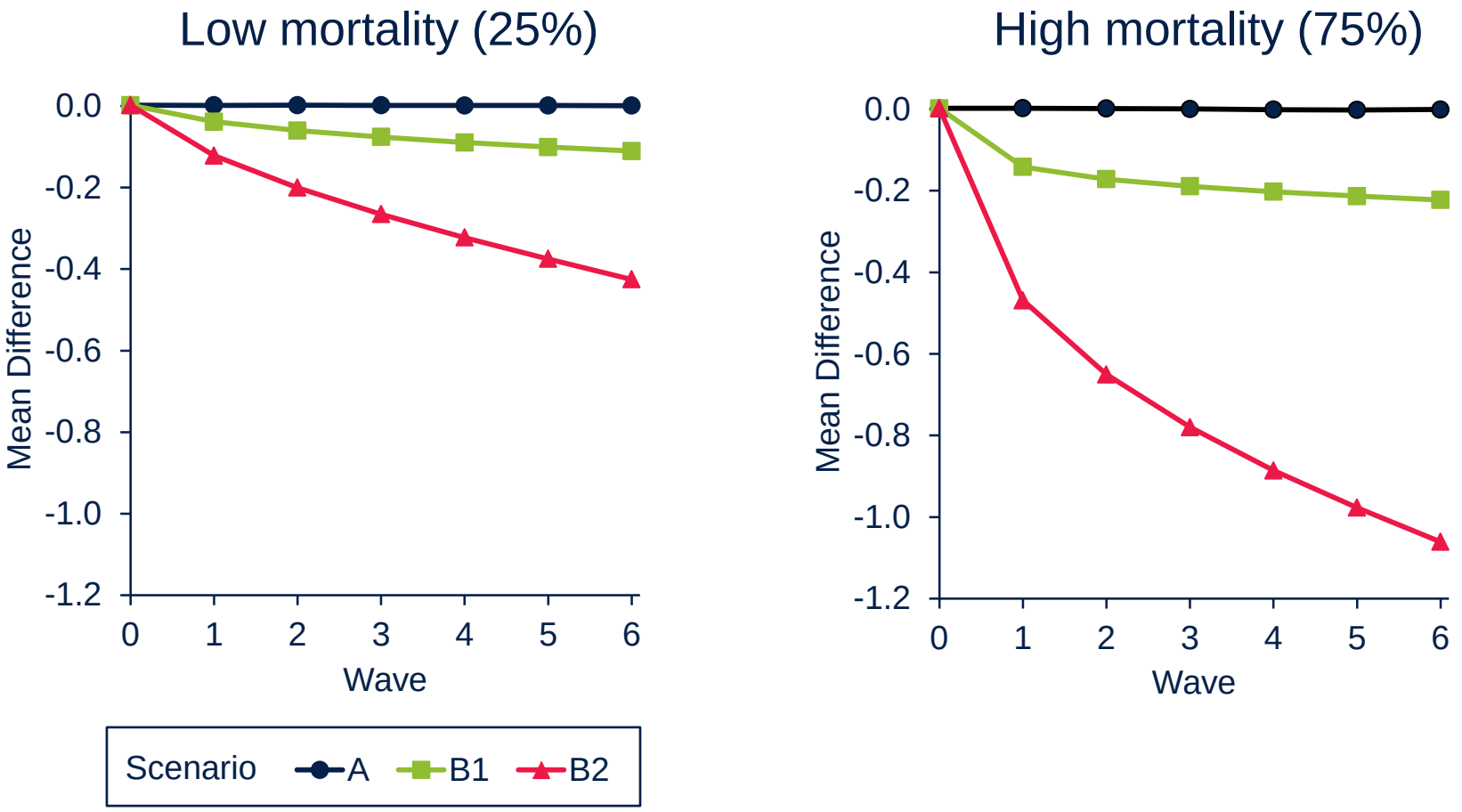


Scenario C exposure and rate of cognitive change affect mortality



A simulation platform to quantify survival bias

- Difference in mean value of U between exposed and unexposed participants alive at each study wave



A simulation platform to quantify survival bias

- Results from B=1,000 simulations on n=1,500 people with up to 7 cognitive assessments over 9 years
- β (true effect of exposure on rate of cognitive decline) = 0

Simulation results for estimated effect of exposure on rate of cognitive change per 10 yrs

	Low mortality scenario (25%)			High mortality scenario (75%)		
	$\bar{\hat{\beta}}$	RMSE	Coverage	$\bar{\hat{\beta}}$	RMSE	Coverage
Scenario A	0.002	0.085	0.954	-0.002	0.120	0.957
Scenario B1	0.048	0.098	0.914	0.094	0.155	0.883
Scenario B2	0.164	0.186	0.527	0.412	0.429	0.081
Scenario C	0.094	0.126	0.809	0.316	0.335	0.229

$\bar{\hat{\beta}}$ = mean of the estimate of interest over all simulations, RMSE = root mean square error (measure of accuracy), Coverage = proportion of times the 95% confidence interval for $\hat{\beta}$ includes β

A simulation platform to quantify survival bias

- Magnitude of bias is strongly influenced by mortality rate
- In high mortality scenarios, substantial bias is easily induced when mortality is influenced by the exposure and cognitive decline
- When mortality is influenced by the exposure and a determinant of cognitive decline, substantial bias only arises when a strong interaction on the hazard ratio scale between the exposure and the determinant of cognitive decline is present

Conclusions

- Selection bias has many forms—can arise from different sources in different study designs
- DAGs can help identify potential selection bias
- Selection of appropriate tools to address selection bias depends on counterfactual contrast of interest
- Sensitivity analyses and simulation studies can quantify potential magnitude of selection bias