Who's in and who's out? Selection bias and potential solutions: Applied examples from aging research

Elizabeth Rose Mayeda, PhD, MPH
Associate Professor
Department of Epidemiology
UCLA Fielding School of Public Health

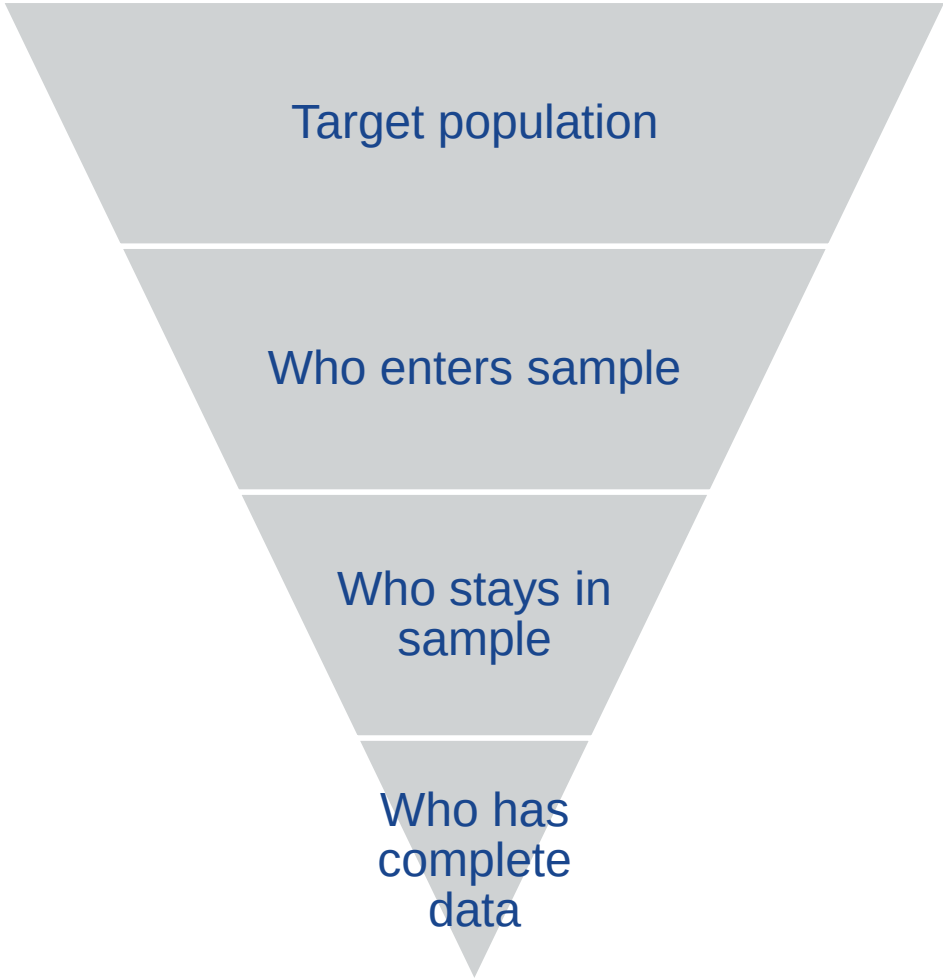
UC Davis Center for Healthcare Policy and Research
October 1, 2024

Selection bias

- Selection bias: Any deviation between the target estimand (i.e., parameter of interest in the target population) and the expected value of the estimate in the sample that arises due to the processes by which observations are included in the sample
- Selection bias can affect internal and/or external validity
- Modern frameworks for selection bias emphasize two phenomena:
  - Restricting to one or more levels of a collider (“collider stratification bias”), internal validity threat
  - Restricting to one or more levels of determinants of the outcome (“generalizability bias”), external validity threat

What gives rise to selection bias?

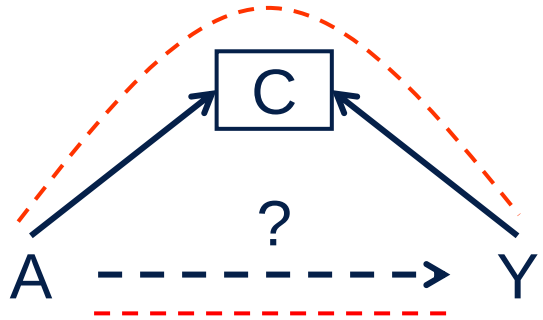
- Arises from differential:
  - Selection into study (by design or self-selection by participants)
    - Who is invited to participate in the study?
    - Who agrees to participate in the study?
    - Who survives?
  - Selective attrition (drop out/death) of enrolled participants
  - Nonresponse, missing data



Collider stratification bias: Restricting to one or more levels of a collider

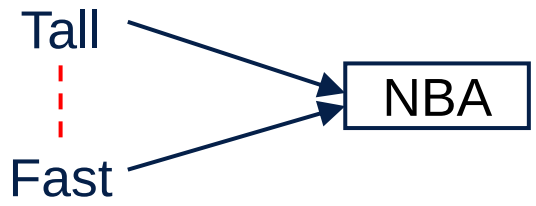
Structure of collider stratification bias

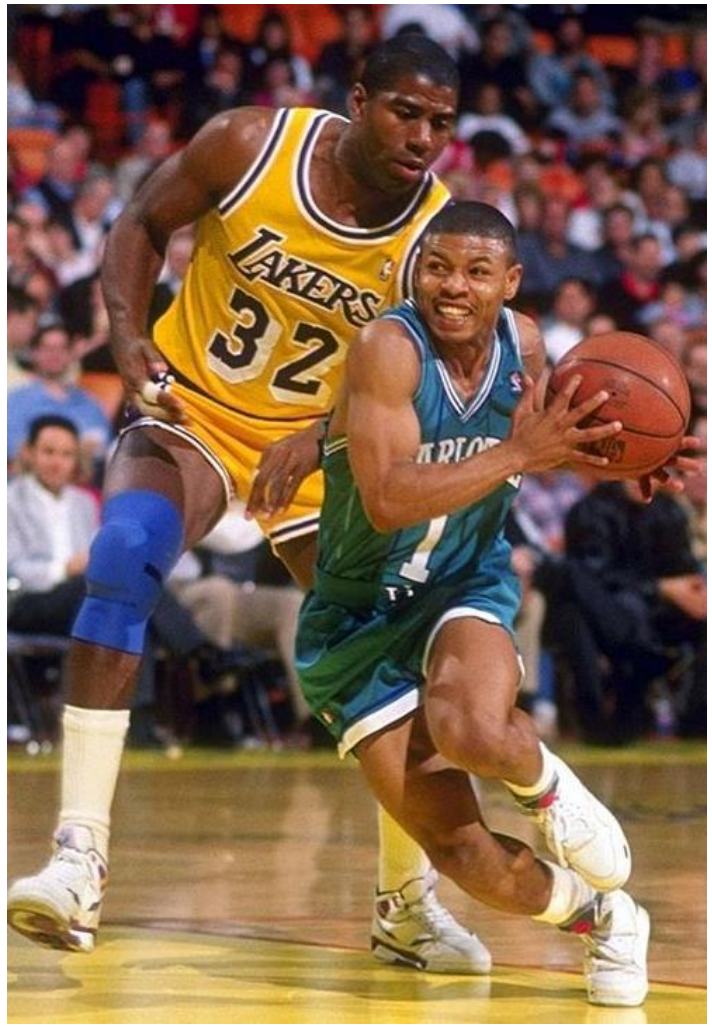
- Conditioning on a common effect of exposure and outcome results in two sources of association between exposure and outcome:
  1. Causal effect of A on Y: A -> Y
  2. Spurious association between A and Y induced by conditioning on collider C: A-> C <- Y



Collider anecdote

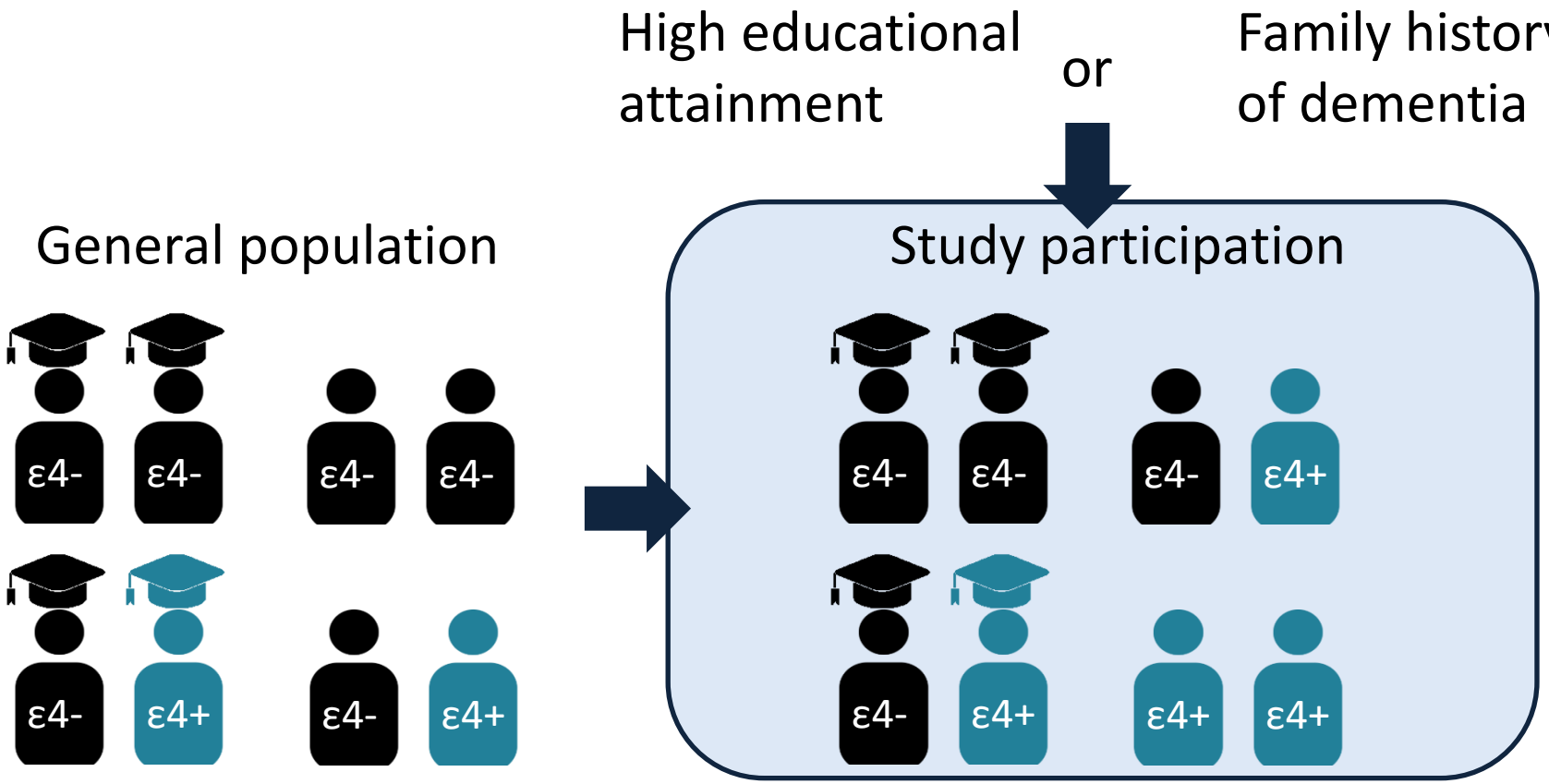
- Some tall people are fast, some are slow
- Some short people are fast, some are slow
- Knowing that a person in the general population is short does not give you information about their speed
- NBA players must be either very tall or very fast
- If you know an NBA player is short …what do you know about his speed





Glymour, MM, 2006. Using causal diagrams to understand common problems in social epidemiology. *Methods in Social Epidemiology*, pp.393-428.

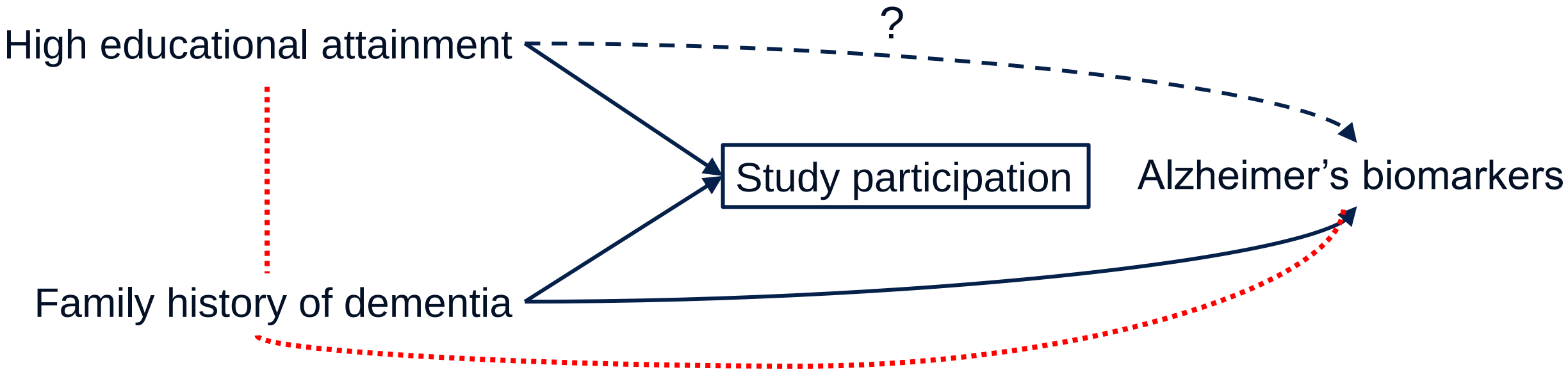
Estimating educational inequalities in Alzheimer's biomarkers from a convenience sample



➔Study conclusions about educational disparities in Alzheimer's biomarkers?

Estimating educational inequalities in Alzheimer's biomarkers from a convenience sample

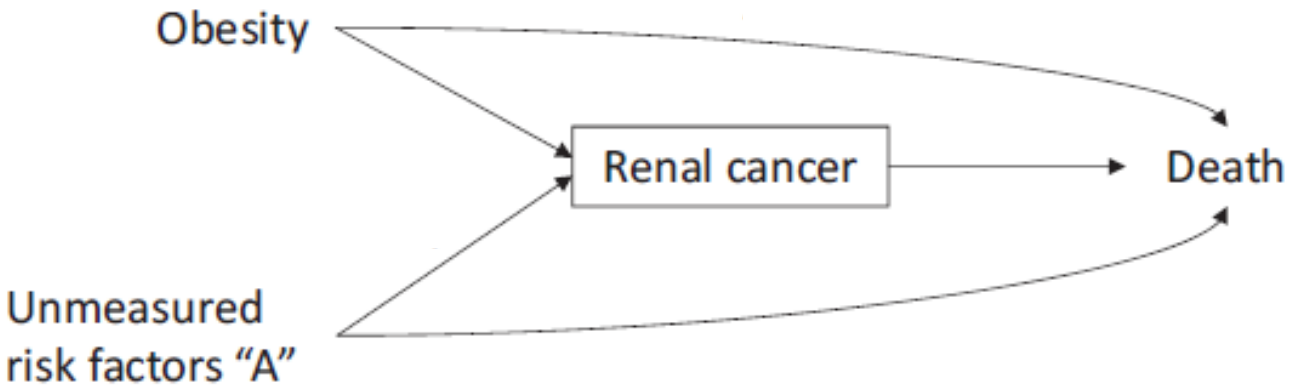
- Directed acyclic graph (DAG) representation of selection process



- Association between education and Alzheimer's biomarkers in the study sample will be a mix of (1) causal effects of education on Alzheimer's biomarkers and (2) spurious association induced by the selection process

“Obesity paradox”

- Harmful effects of obesity on mortality in the general population
- Many studies report an “obesity paradox”
  - Among those with chronic disease (CVD, cancer), better survival among individuals with overweight/obese BMIs vs. BMIs in “normal weight” range
- Does this reflect a causal effect or a spurious (non-causal) association?

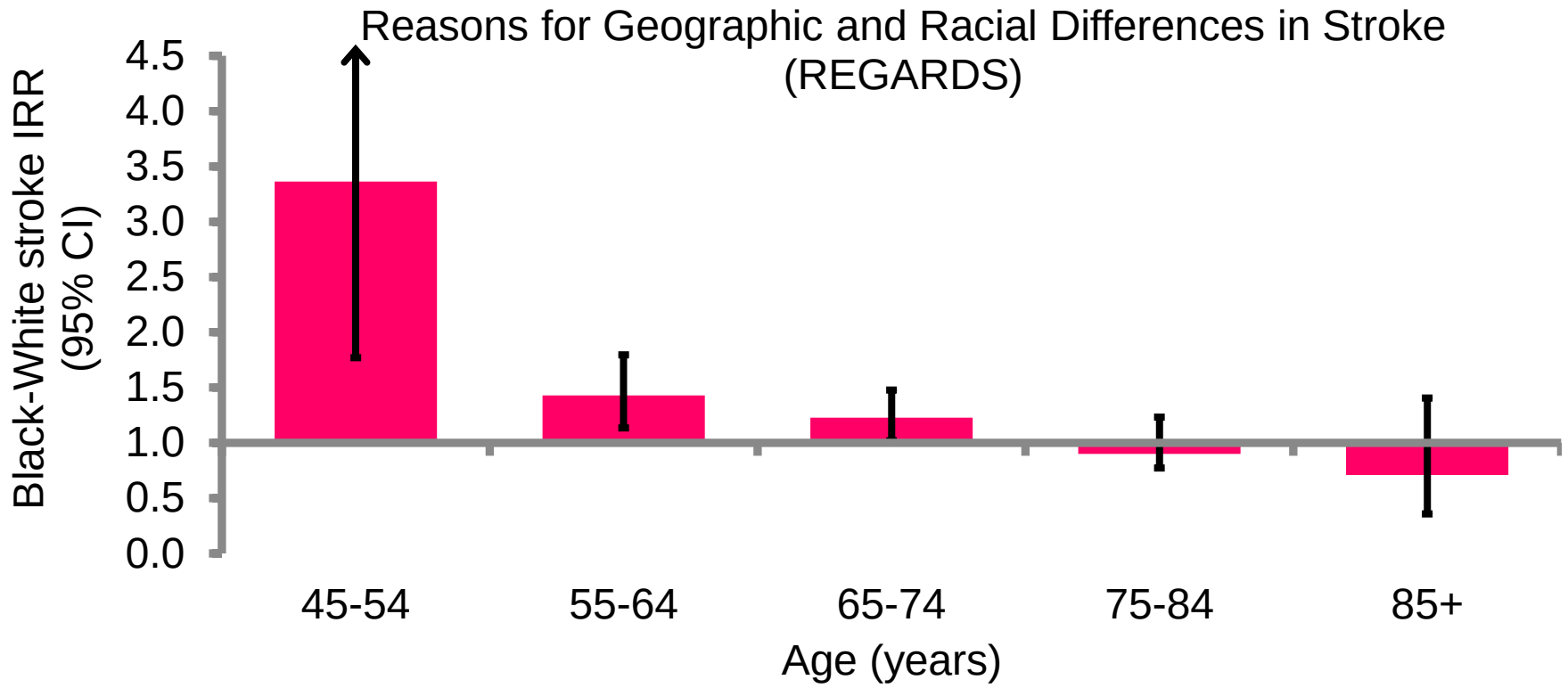


Mayeda, *Cancer Epidemiol. Biomark. Prev* 2017

Example: Can Survival Bias Explain the Age Attenuation of Racial Inequalities in Stroke Incidence?

Racial inequalities in stroke

- Stroke is a leading cause of disability and death in the U.S.
- Qualitative change in racial inequalities in stroke incidence between middle and late life



Benjamin *et al., Circulation* 2017;
Personal communication with Dr. George Howard, REGARDS Study PI, December 2016

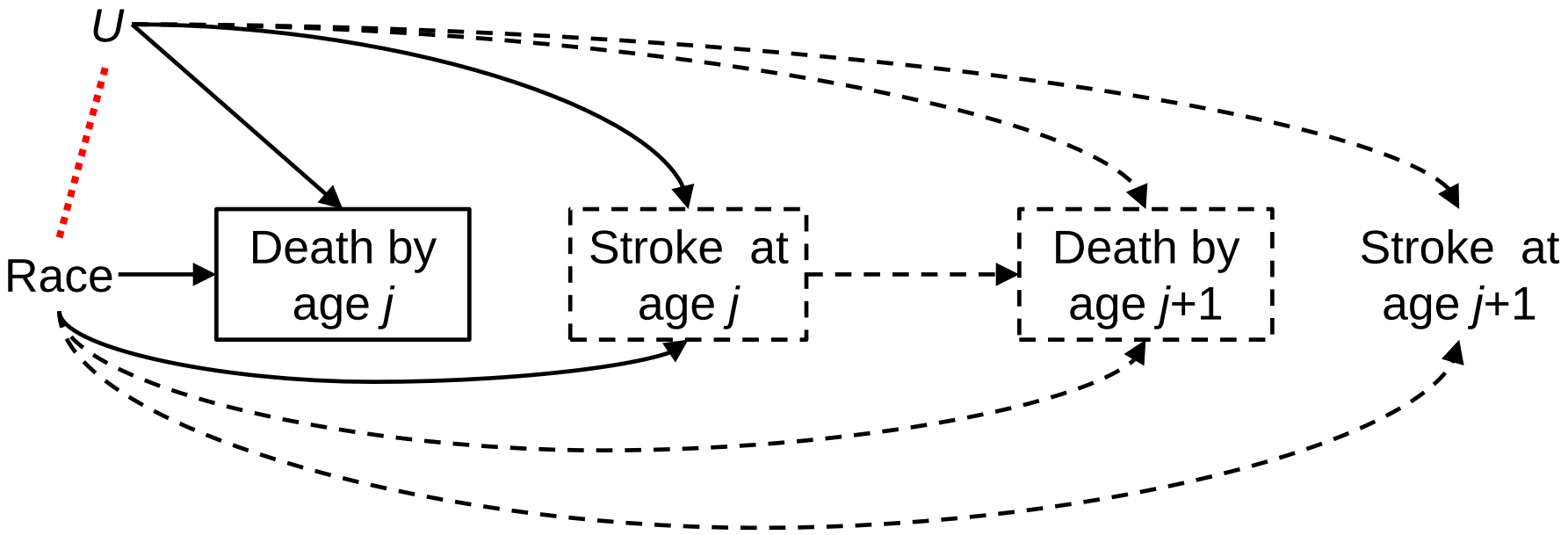
UCLA

What is driving the age-attenuation of racial inequalities in stroke?

- Causal explanation: Improved social conditions for Black Americans at older ages
- Selective survival: Among survivors to old age, Black Americans represent a more selected, healthier population than White Americans
    - Median survival for 1919-1921 birth cohort: 65 years White Americans, 50 years Black Americans

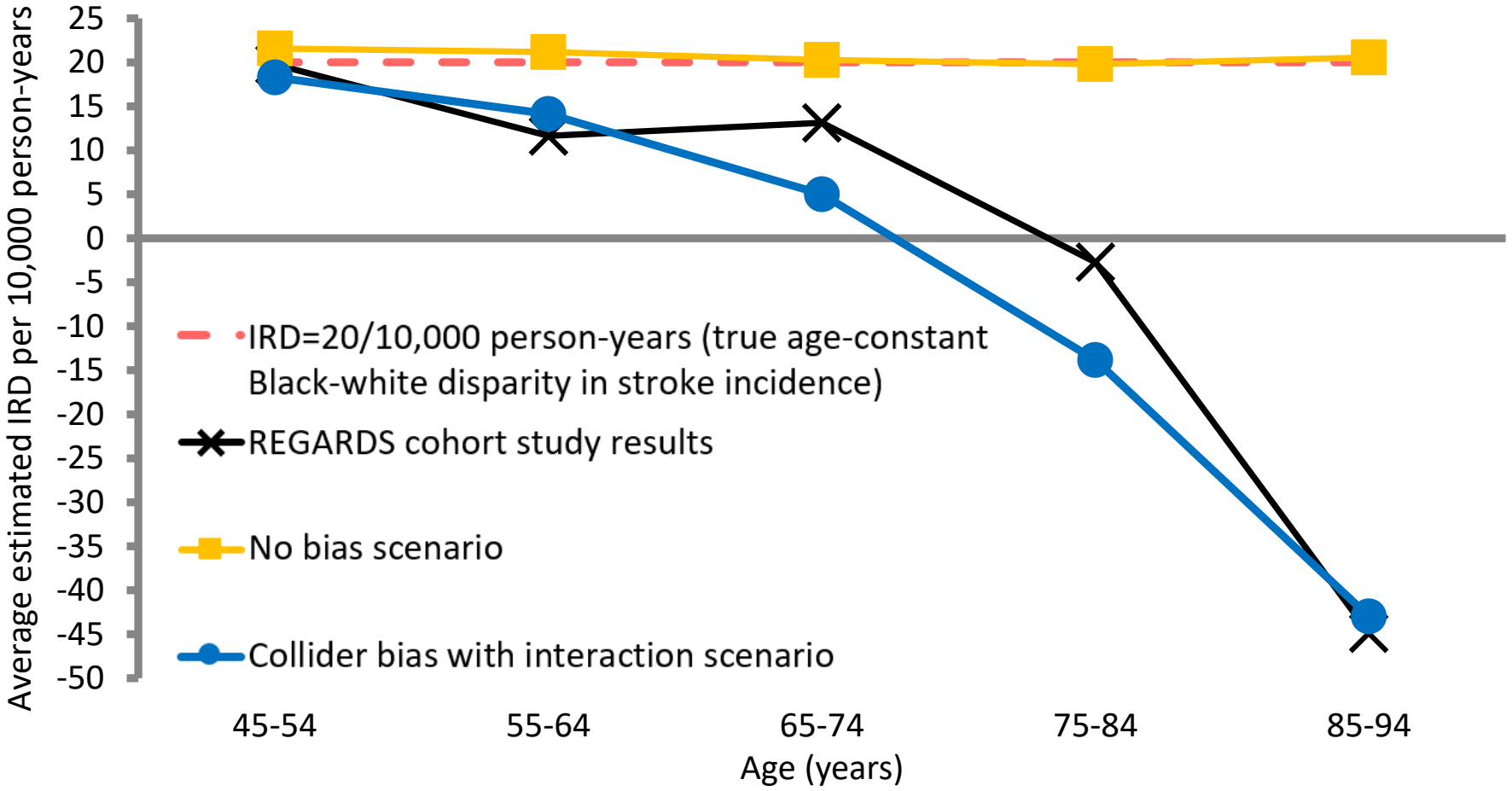
Causal scenarios

- **Collider bias scenario:** *U* directly influences stroke risk and mortality risk (HR_{stroke}=1.5; $HR_{mortality}$=1.5)
- **Collider bias with interaction scenario:** *U* directly influences stroke risk and mortality risk for Blacks; *U* has no direct effect on mortality for Whites



Understanding changes in health inequalities across the lifecourse: racial inequalities in stroke incidence

Average observed Black-white stroke incidence rate difference (IRD) by age band (2,000 simulated samples)



UCLA

Investigating and Remediating Selection Bias in Geriatrics Research: The Selection Bias Toolkit

Hailey R. Banack, PhD, Jay S. Kaufman, PhD,† Jean Wactawski-Wende, PhD,* Bruce R. Troen, MD,‡ and Steven D. Stovitz, MD, MS§*

Table 1. Selection Bias Toolkit

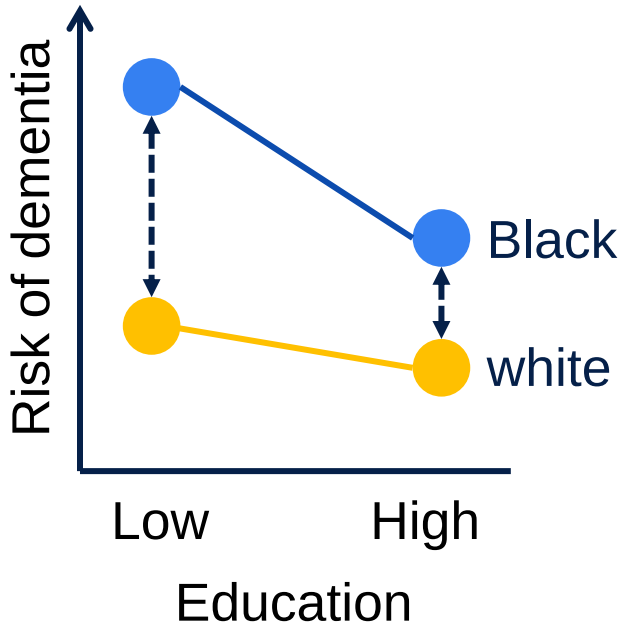
Study stage	Method	Description
Design	Causal diagrams	Graphic representation of causal effects between variables. Useful for identifying sources of selection bias and describing assumptions about relationships between variables of interest.
	Data collection	Minimizing barriers to completing data collection in geriatric setting. Collecting as much information as possible on potential predictors of censoring.
Analysis	Descriptive statistics	Examine distribution of important variables among individuals who remained in the study compared with those who were lost to follow-up or died during the follow-up period.
	Model determinants of selection	Use a regression model (ie, logistic regression) to determine which variables are predictors of censoring. In certain situations it is possible to include predictors of censoring in the main analytic models.
	Sensitivity analysis	Supplementary analyses that involve hypothesizing about a range of different selection mechanisms and estimating the effect of specific selection scenarios on study results.
	Bias analysis	Similar to sensitivity analyses, bias analysis includes hypothesizing about the magnitude of specific selection mechanisms (sampling fractions) and adjusting effect estimates by dividing the biased effect estimate by the selection bias factor.
	Inverse probability weighting	Used to correct for selection bias by up-weighting individuals who remain in the study cohort to account for themselves as well as those with similar characteristics who have been lost to follow-up.
	Principal stratification	The effect of exposure on outcome in the subset of individuals who would have survived to old age, regardless of exposure status.
	Multiple imputation	Replacing missing values (for individuals lost to follow-up) with a set of plausible variables to create a complete simulated data set.
	Instrumental variables	Method to control for effect of unmeasured selection factors in observational studies; includes selecting an instrument that mimics the random group allocation of a randomized study. Dependent on having an appropriate and valid instrument.

J Am Geriatr Soc. 2019 Jun 18. doi: 10.1111/jgs.16022.

Generalizability bias: Restricting to one or more levels of determinants of the outcome

Generalizability bias example

- People with high educational attainment are overrepresented in studies compared to people with low educational attainment
- If racial disparities in dementia are wider among those with low vs. high educational attainment, a study of mostly highly-educated participants will show little evidence of racial disparities in dementia
- Results from a study of mostly highly-educated participants would not generalize to the U.S. population of older adults (which includes a greater mix of people with high and low educational attainment)

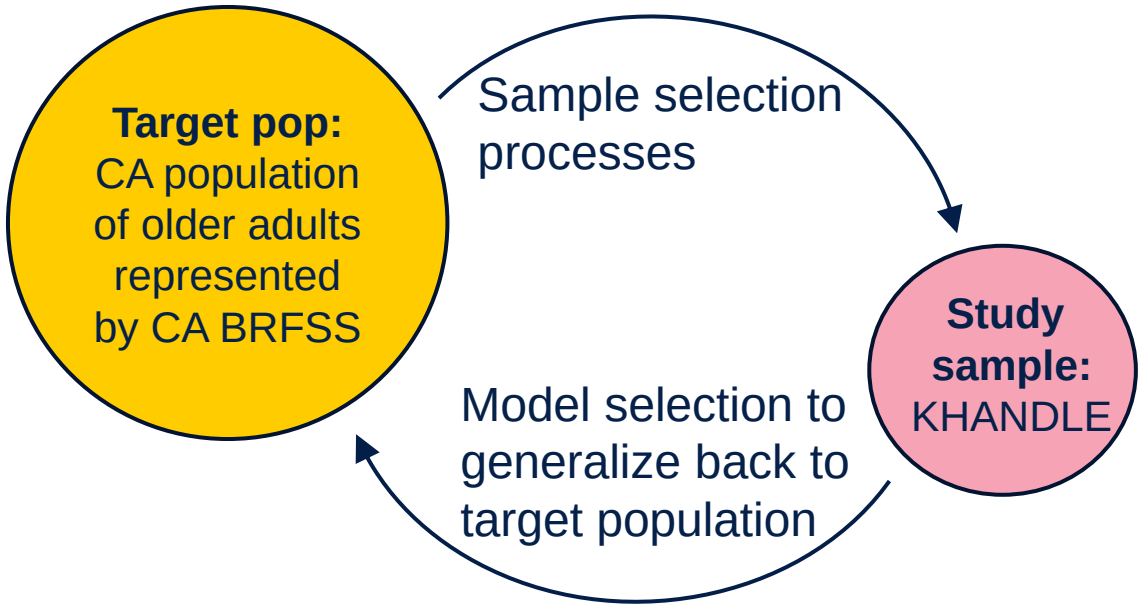


Trial samples are often highly selected

- Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial evaluated the effects of intensive glycemic control on cardiovascular events among people age ≥ 40 years with type 2 diabetes at high cardiovascular risk
- We applied ACCORD eligibility criteria to a nationally representative sample of adults age ≥ 40 years with diabetes in the National Health and Nutrition Examination Survey (NHANES)
  - ACCORD represented a minority of U.S. middle-aged and older adults with diabetes
    - 12% met eligibility criteria when using HbA1c ≥ 6% to define the target population
    - 39% met eligibility criteria when using HbA1c ≥ 7.5% to define the target population

Efforts to improve knowledge about cognitive impairment in the general population

- KHANDLE is more diverse than many studies, but selection issues persist
- Transportability methods to extend findings from KHANDLE to CA population



	KHANDLE N=1,708	CA BRFSS N=12,399
Age, mean (SD)	76.0 (6.8)	73.9 (6.9)
Male, %	40.5	44.1
Race/ethnicity, %		
Asian	24.3	14.0
Black	25.9	6.0
Latino	20.4	18.6
White	29.4	61.4
Educational attainment, %		
0-8 years	3.2	10.0
Some high school	3.8	6.1
High school diploma/GED	9.9	18.0
Trade school/technical school/some college	35.0	34.9
College graduate or higher	48.1	30.9
Self-rated health "good" or better, %	81.3	75.3

Data and primary outcome

- Study sample: KHANDLE
- Target population: CA BRFSS (2014-2018, weighted to be CA-representative)
  - Age 65+
  - Asian, Black, Latino, and White individuals
  - Insured
  - English- and Spanish-speaking
  - Free of dementia diagnosis
- Primary outcome (available only in KHANDLE): Cognitive impairment

Candidate variables for weights

- Variables that are associated with cognitive impairment and differ in distribution in KHANDLE & CA BRFSS
- Limited to variables available for harmonization in KHANDLE & CA BRFSS
- Sociodemographic characteristics:
  - Age
  - Sex
  - Marital status
  - Educational attainment
  - Per-capita income
  - Interview language (Spanish/English)
- Health characteristics:
  - Self-rated health
  - Smoking status
  - Vision impairment
  - Physical activity
  - Activities of daily living (walking/climbing stairs and dressing/bathing)

Approach

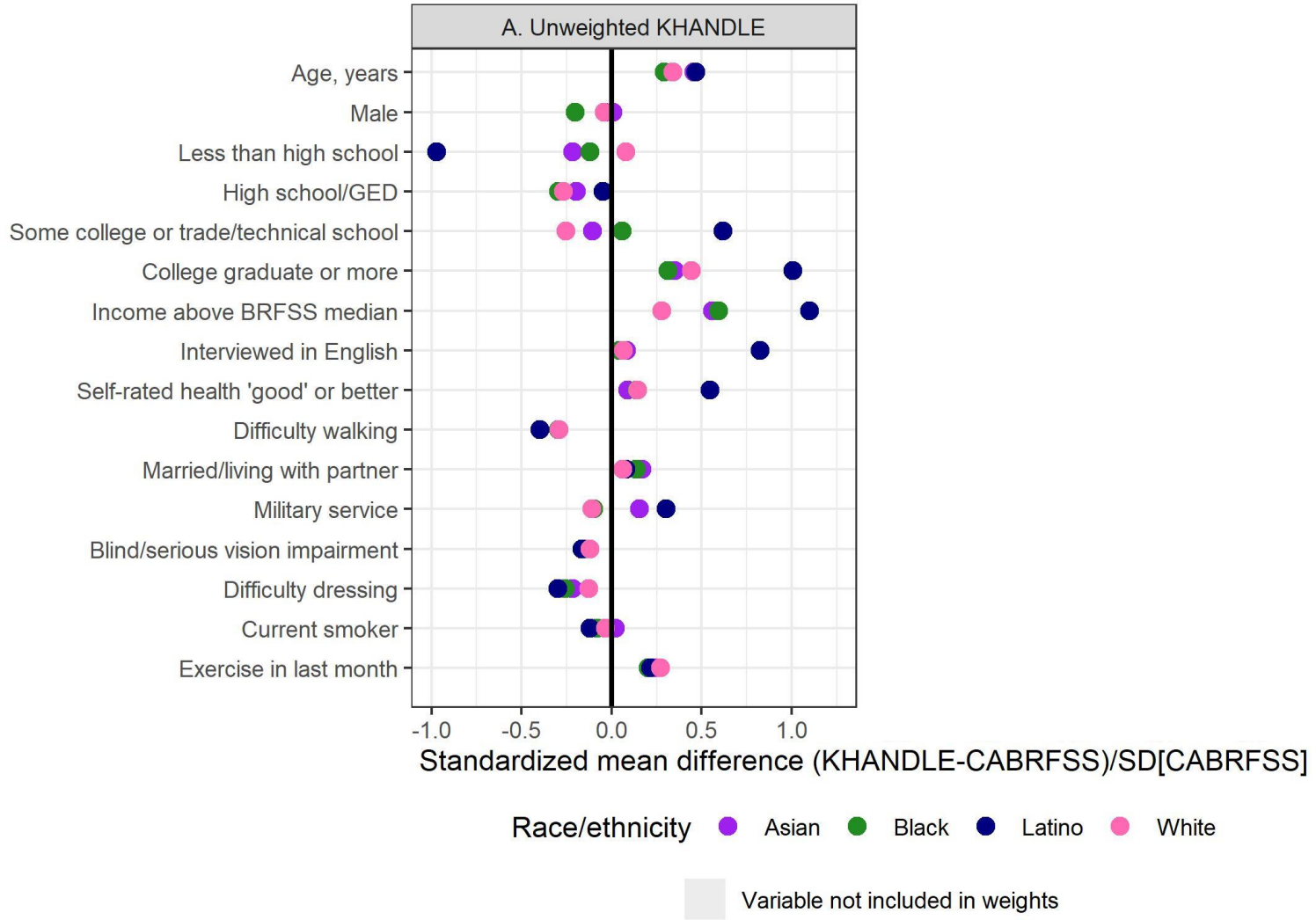
- Pooled KHANDLE and CA BRFSS datasets
- Assessed covariate balance in KHANDLE and CA BRFSS with standardized mean differences (<0.25 deemed adequate balance)
- Iteratively developed inverse odds of selection weights using logistic regression to estimate probability of KHANDLE participation:

$$W_i = [P(S_i=0|\mathbf{Z_i})/P(S_i=1|\mathbf{Z_i})]\times[P(S_i=1)/P(S_i=0)]$$

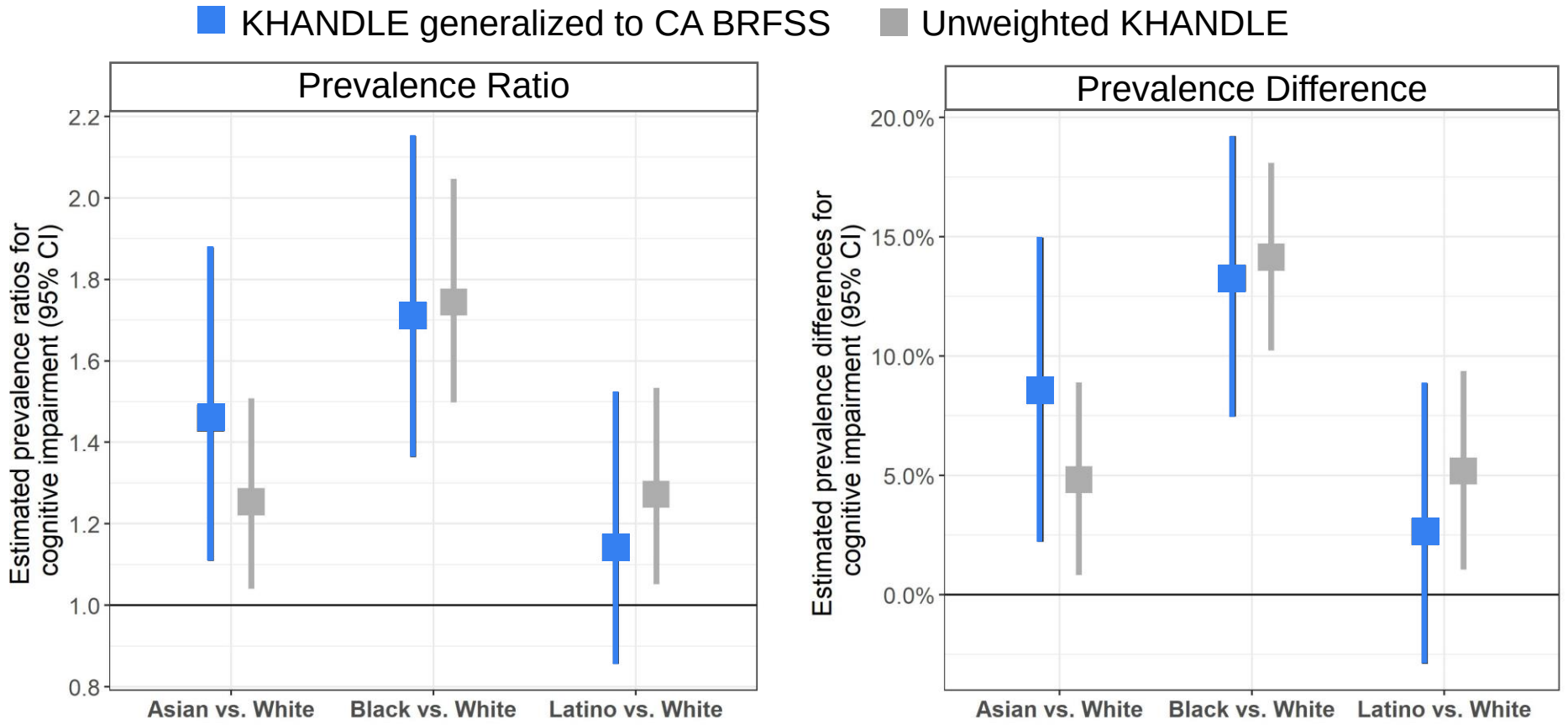
- Assessed covariate balance in KHANDLE and CA BRFSS after applying weights to KHANDLE
- Applied weights to KHANDLE to estimate racial/ethnic inequalities (prevalence ratios and differences) in p(cognitive impairment) in target population

Covariate balance (race/ethnicity-specific standardized mean differences) between KHANDLE CA-BRFSS

- **Before weighting:** KHANDLE was older, had more education, higher income, and was slightly healthier than CA-BRFSS



Estimates of racial/ethnic inequalities in cognitive impairmen prevalence: KHANDLE generalized to CA population of older adults

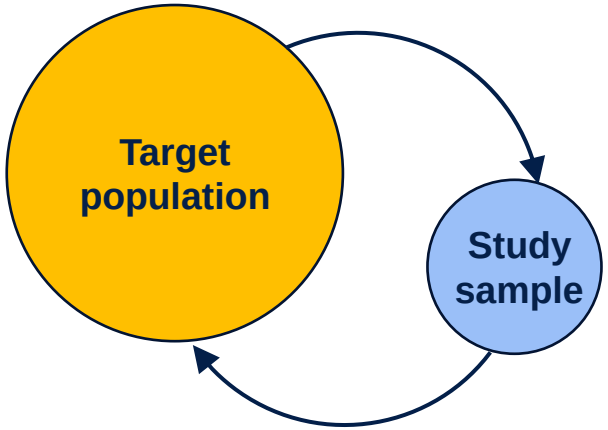


 Hayes-Larson, *Alzheimer's Dement* 2022

UCLA

What can we do to enhance generalizability of research findings?

- Plan studies with generalizability in mind
  - Consider the target population during the study design phase
    - Defining eligibility criteria
    - Developing recruitment and retention plans
- Once data are collected:
  - Rigorously compare study samples to target population
  - Apply statistical tools to extend study findings to target populations (Hayes-Larson, *Epidemiology* 2024)
  - Evaluate evidence with a critical eye on generalizability



Acknowledgements

- Eleanor Hayes-Larson
- Taylor Mobley
- Yingyan Wu
- Joey Fong
- Juliet Zhou
- Ryo Ikesu
- Paloma Rojas Saunero
- Natalie Gradwohl
- Adiba Hassan
- Gina Nam
- Layla Ruiz
- Maria Glymour
- Rachel Whitmer
- Dan Mungas
- Paola Gilsanz
- Gil Gee
- Ron Brookmeyer
- Marissa Seamans
- Daniel Westreich
- Hailey Banack
- Kirsten Bibbins-Domingo
- Adina Zeki Al Hazzouri
- MELODEM (NIA R13AG064971)
- ΨMCA (NIA R13AG030995)

- R00AG053410 (Mayeda), R01AG063969 (Mayeda), R56AG069126 (Mayeda), R01AG6052132 (Whitmer, Mayeda, Glymour, Gilsanz), R01AG074359 (Casey & Mayeda)

Thank you!

mayeda@ucla.edu