Second-guessing selection bias

Re-evaluating our approaches

Louisa Smith

LVX VERITAS VIRTVS
Northeastern University
Bouvé College of Health Sciences

The Roux Institute
Northeastern University

What is selection bias?

Accepted version to appear in *American Journal of Epidemiology*

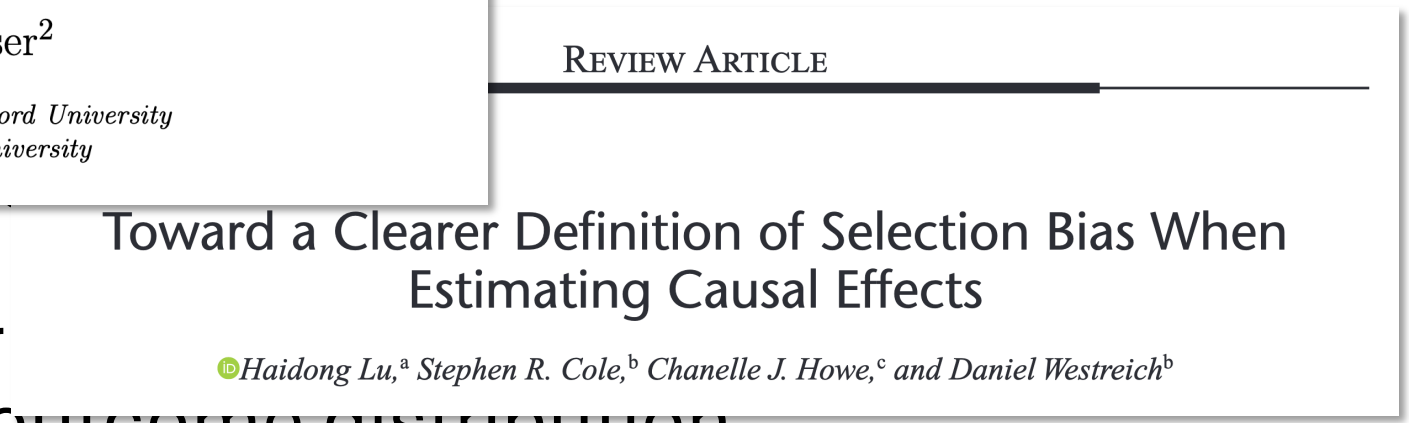
Simple graphical rules to assess selection bias in general-population and selected-sample treatment effects

Maya B. Mathur*[1] and Ilya Shpitser[2]

[1] *Quantitative Sciences Unit, Department of Medicine, Stanford University*
[2] *Department of Computer Science, Johns Hopkins University*

Type 2
selection bias

effect in the target



REVIEW ARTICLE

Toward a Clearer Definition of Selection Bias When Estimating Causal Effects

Haidong Lu,[a] Stephen R. Cole,[b] Chanelle J. Howe,[c] and Daniel Westreich[b]

Conditioning on a collider
well as effect modifier or outcome distribution

- Measured? Maybe can fix, maybe can't
- Unmeasured? Might find an association where there's not one

Type1
selection bias

What is selection bias?

When researchers only choose certain types of people to participate in research, like those who are healthier than average

Inequity in participation (without reference to an estimand)

Type 3 selection bias?

Selection bias of all types in the context of the *All of Us* Research Program

NIH-funded study attempting to recruit 1 million Americans

No sampling strategy – volunteer recruitment

Targeted recruitment of communities previously underrepresented in biomedical research

Representation

Ancestry:

- **Race:** People who select a single race other than White (e.g., Asian), or who select more than one race
- **Ethnicity:** People who select an ethnicity other than those listed under the race of White (e.g., Japanese)

Age: People who are 65 years of age or older at the time of primary consent

Sexual and gender minorities:

- **Sex assigned at birth:** People who self-report intersex as their sex at birth
- **Sexual orientation:** People who select any sexual orientation choice other than straight (e.g., gay, lesbian, bisexual, queer, asexual, etc.)
- **Gender identity:** People who select any gender identity choice other than man or woman (e.g., non-binary, transgender, genderfluid, questioning, etc.)

Income: People with an annual household income at or below 200% of the Federal Poverty Level (FPL) based on residency (defined as the 48 contiguous states, Alaska, or Hawaii) and household size

Educational attainment: People without a high school diploma or GED

Geography: Residents of established rural and non-metropolitan zip codes, based on the HRSA Federal Office of Rural Health Policy data files

Disability: People with a physical, functional, cognitive, or other condition that substantially limits one or more life activities

Healthcare Access & Utilization: People with inadequate access to healthcare who lack health insurance, have no source of primary care, or who are unable to obtain needed medical care within the past 12 months due to selected barriers

https://www.researchallofus.org/frequently-asked-questions/

Representation

Allows us to ask questions we otherwise couldn't

Middle Eastern / North African not captured in vital statistics

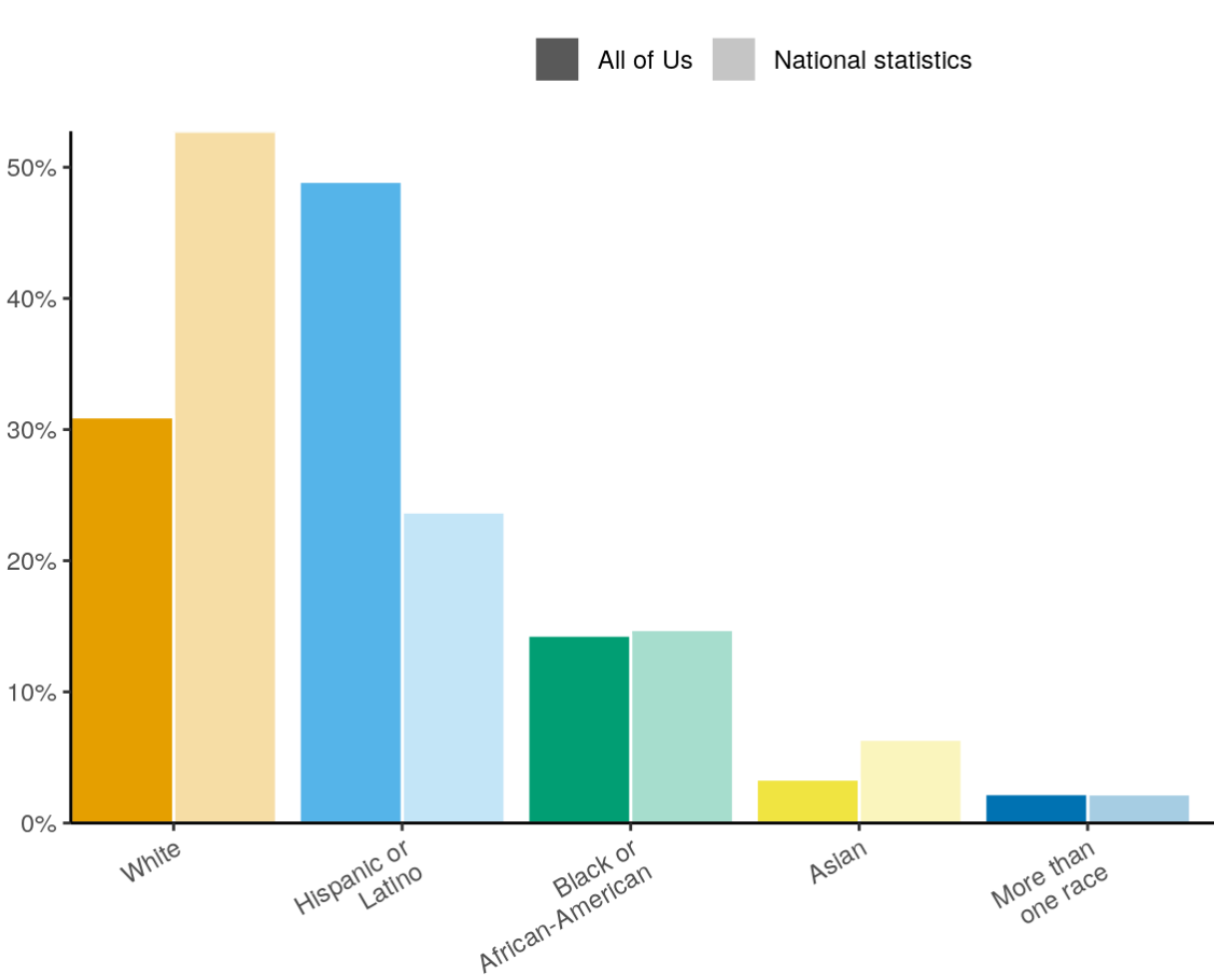
Gender identity and sexual orientation not captured in vital statistics

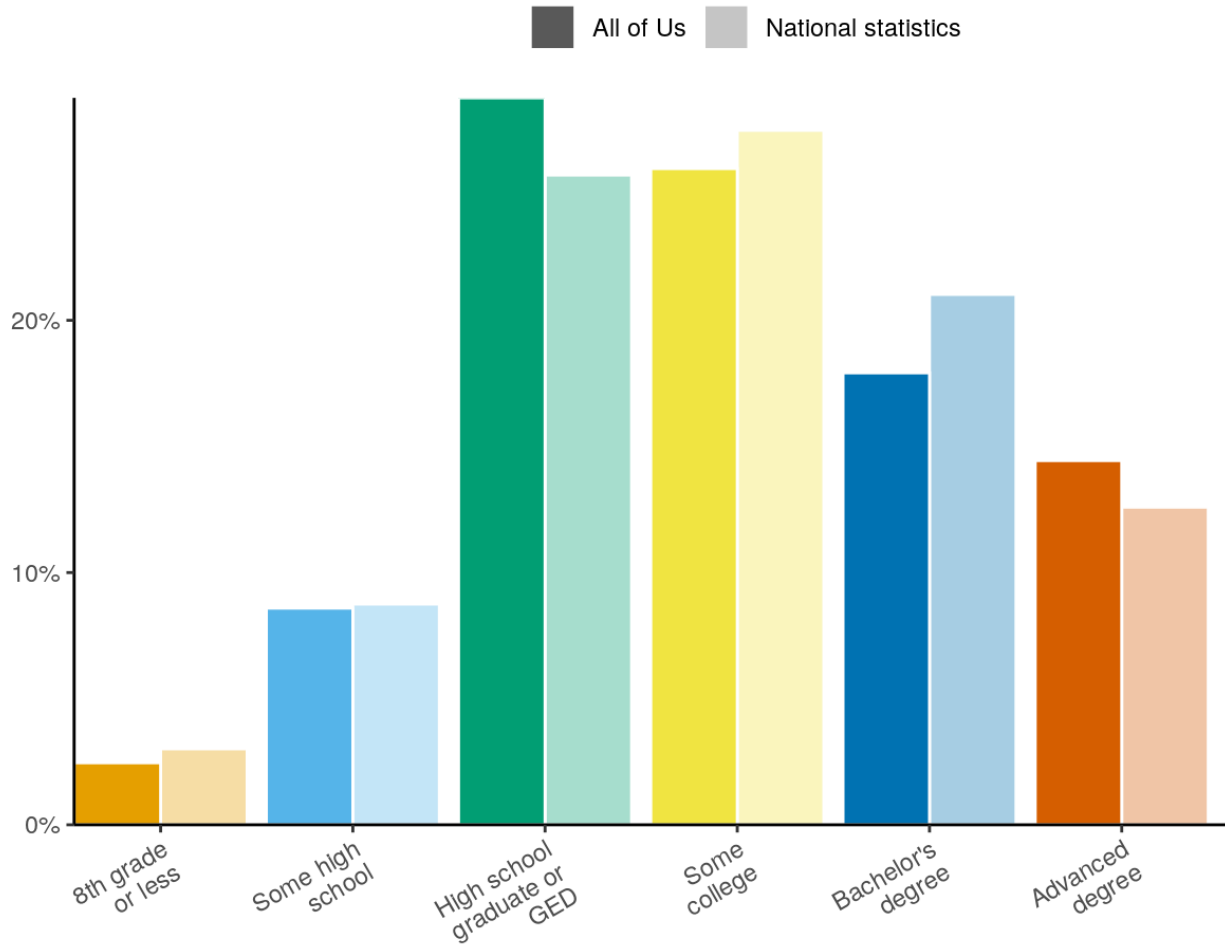
Smith et al, MedRxiv 2024

Pregnant people in All of Us	N = 14,237
Race/ethnicity	
Hispanic or Latino	6,044 (43.2%)
White	4,702 (33.6%)
Black or African-American	2,244 (16.0%)
Asian	465 (3.3%)
More than one race	304 (2.2%)
Other	118 (0.8%)
Middle Eastern or North African	95 (0.7%)
Native Hawaiian or Other Pacific Islander	25 (0.2%)
Gender	
Woman	14,018 (99.4%)
Man	28 (0.2%)
Other/multiple	52 (0.4%)
Sexual orientation	
Bisexual	833 (6.0%)
Gay/lesbian	89 (0.6%)
None	195 (1.4%)
Straight	12,795 (92%)

But lack of representativeness

Distribution of demographics of live births in All of Us compared to vital statistics data





Smith et al, MedRxiv 2024

Large-scale volunteer databases

- All of Us (US)
- UK Biobank (UK)
- CanPath (Canada)
- NAKO (Germany)
- Biobank Japan (Japan)
- Taiwan Biobank (Taiwan)
- FinnGEN (Finland)


nature aging

PERSPECTIVE

https://doi.org/10.1038/s43587-022-00277-x

Check for updates

The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging

Carol Brayne[1]✉ and Terrie E. Moffitt[2,3,4]

3000+ articles last year (PubMed)

It’s clear there’s selection happening… is there bias?

- It looks like (in All of Us) we’re avoiding the inequities in selection of previous research

- But can we address other forms of selection bias?

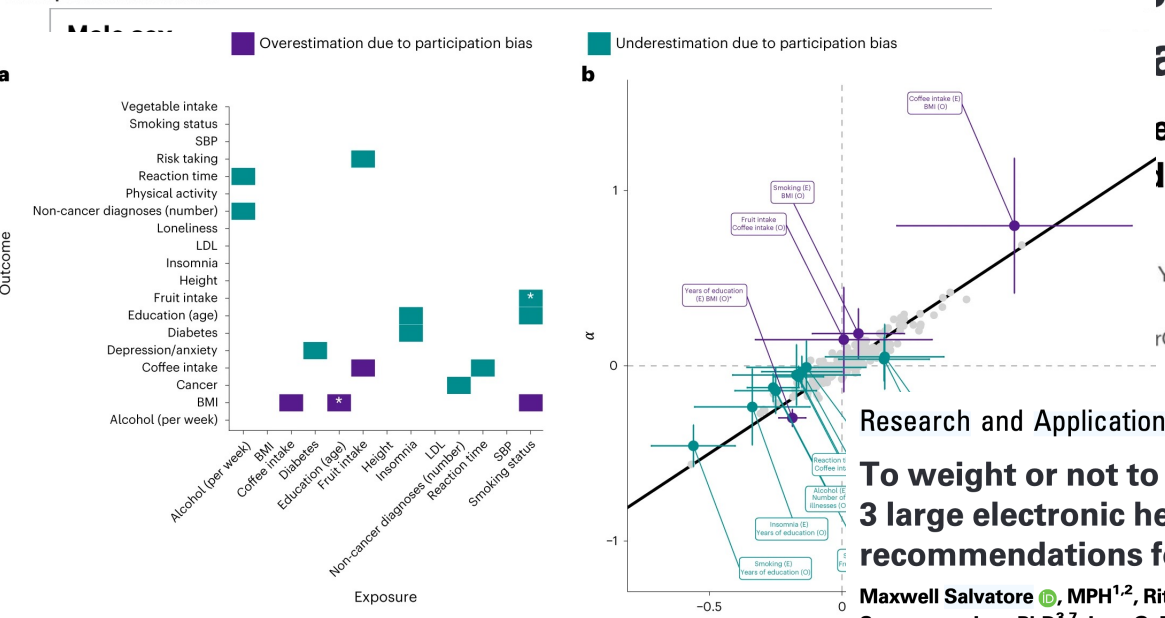
Article https://doi.org/10.1038/s41562-023-01579-9

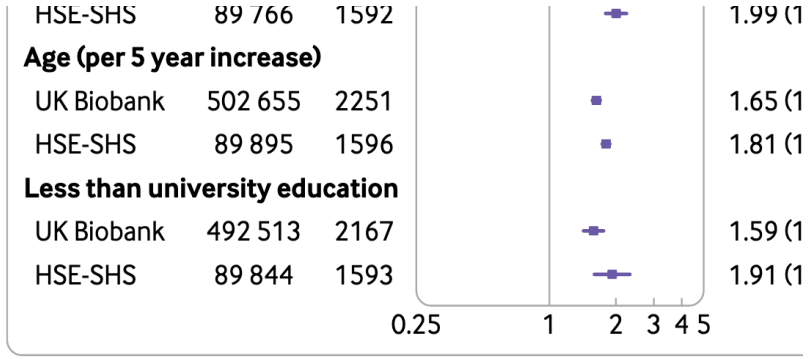
Participation bias in the UK Biobank distorts genetic associations and downstream analyses

Received: 27 September 2022

Accepted: 7 March 2023

Tabea Schoeler[1,2], Doug Speed[3], Eleonora Porcu[4], Nicola Pirastu[5], Jean-Baptiste Pingault[2,6] & Zoltán Kutalik[1,7,8]





ational Journal of Epidemi
//doi.org/10.1093/ije/dyae05
al article

ginal article

eweighting
as due to v

erd van Alten
dries T Marees[1]

YearsEducation ~ Female
rOfCars ~ YearsEducation

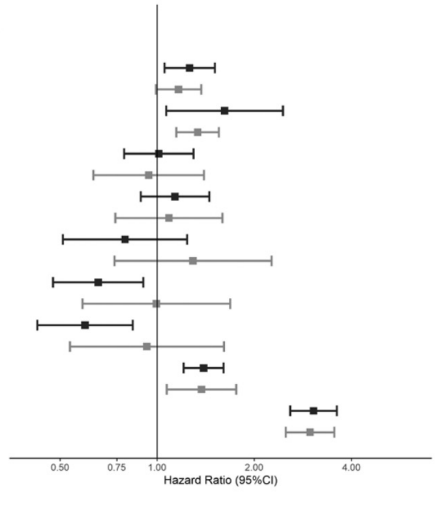
ORIGINAL ARTICLE

OPEN

Is Cohort Representativeness *Passé*? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank

Emmanuel Stamatakis,[a] *Katherine B. Owen,*[b] *Leah Shepherd,*[b] *Bradley Drayton,*[b] *Mark Hamer,*[c] *and Adrian E. Bauman*[b]

Variables	level	Reference	Hazard ratio (95% CI)
Physical activity level	>0, <7.5 MET-hrs/wk	≥7.5 MET-hrs/wk	1.26 (1.05, 1.51)
			1.17 (0.99, 1.37)
	No physical activity		1.62 (1.07, 2.45)
			1.33 (1.15, 1.55)
Fruit and vegetable consumption	5-9 portions/day	≥10 portions/day	1.01 (0.79, 1.30)
			0.94 (0.63, 1.40)
	<5 portions/day		1.14 (0.89, 1.45)
			1.09 (0.74, 1.59)
Alcohol use frequency	Previous	Never	0.79 (0.51, 1.24)
			1.29 (0.74, 2.26)
	Current: <5 times per week		0.66 (0.48, 0.91)
			1.00 (0.59, 1.69)
	Current: ≥5 times per week		0.60 (0.43, 0.84)
			0.93 (0.54, 1.61)
		ver	1.39 (1.21, 1.61)
			1.37 (1.07, 1.76)
			3.05 (2.59, 3.60)
			2.98 (2.50, 3.54)

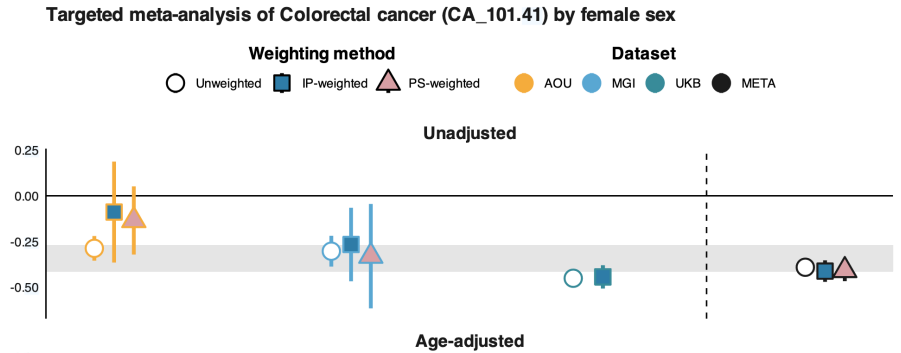


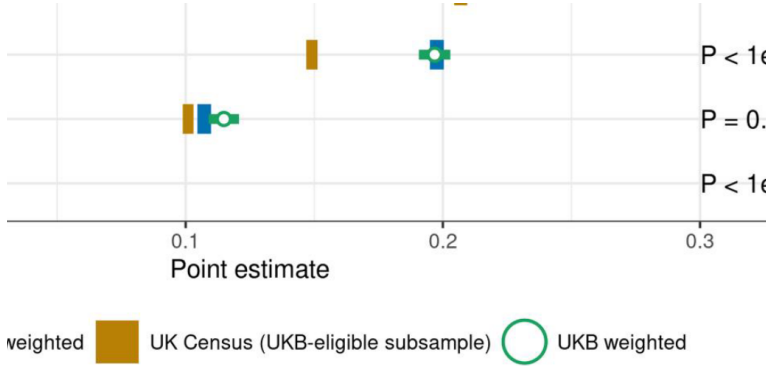
usted for all lifestyle risk factors, and additionally adjusted for age, sex, and highest e
ı lifestyle risk factor for all cardiovascular disease (CVD) mortality, excluding people w
ter exclusion). MET indicates metabolic equivalents.

Research and Applications

To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice

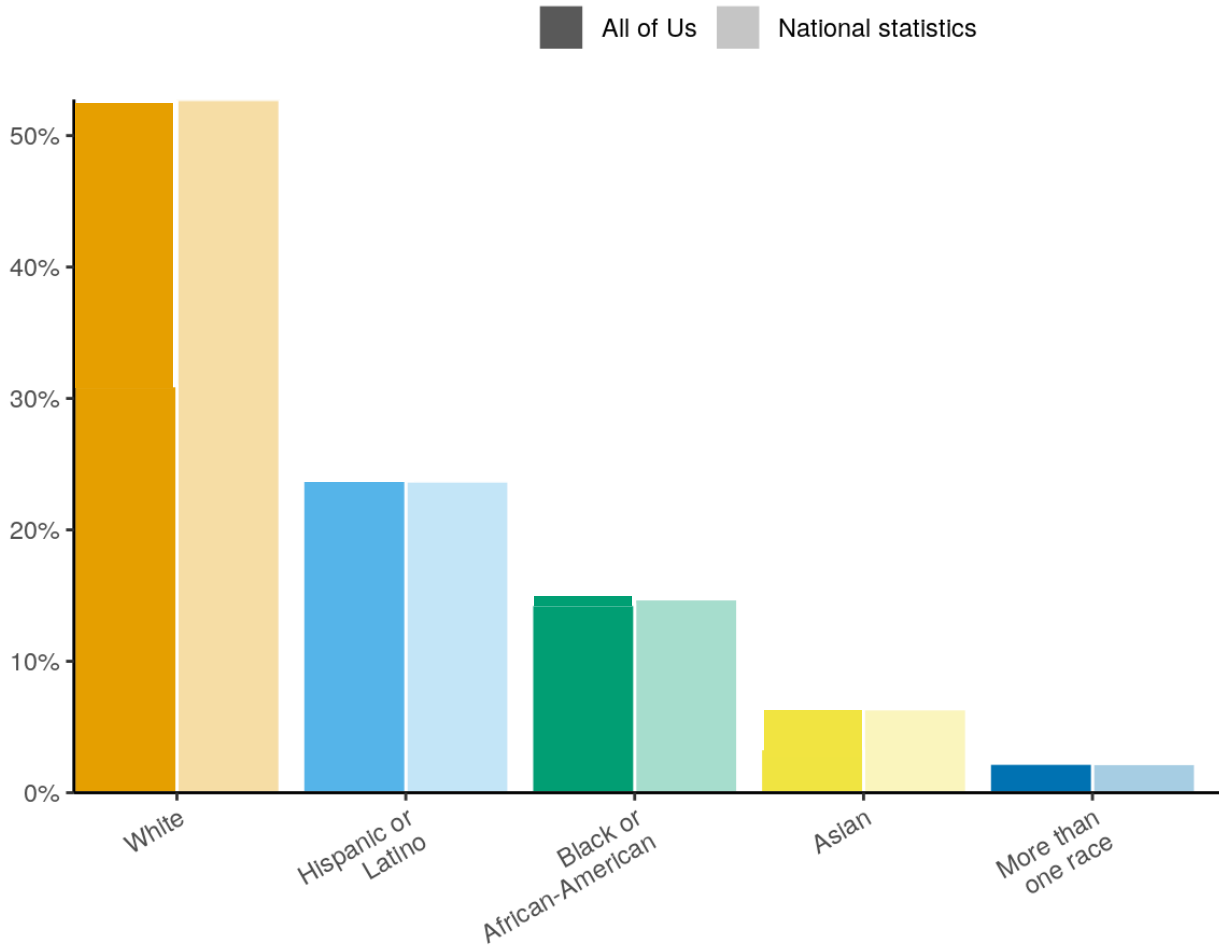
Maxwell Salvatore, MPH[1,2], Ritoban Kundu, MS[2,3], Xu Shi, PhD[3], Christopher R. Friese, PhD[4,5,6], Seunggeun Lee, PhD[3,7], Lars G. Fritsche, PhD[2,3,4], Alison M. Mondul, PhD[1,4], David Hanauer, MD[8], Celeste Leigh Pearce, PhD[1,4], Bhramar Mukherjee, PhD[1,2,3,*]

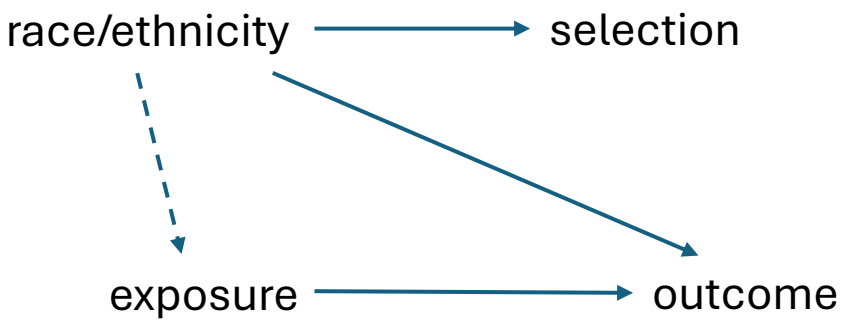




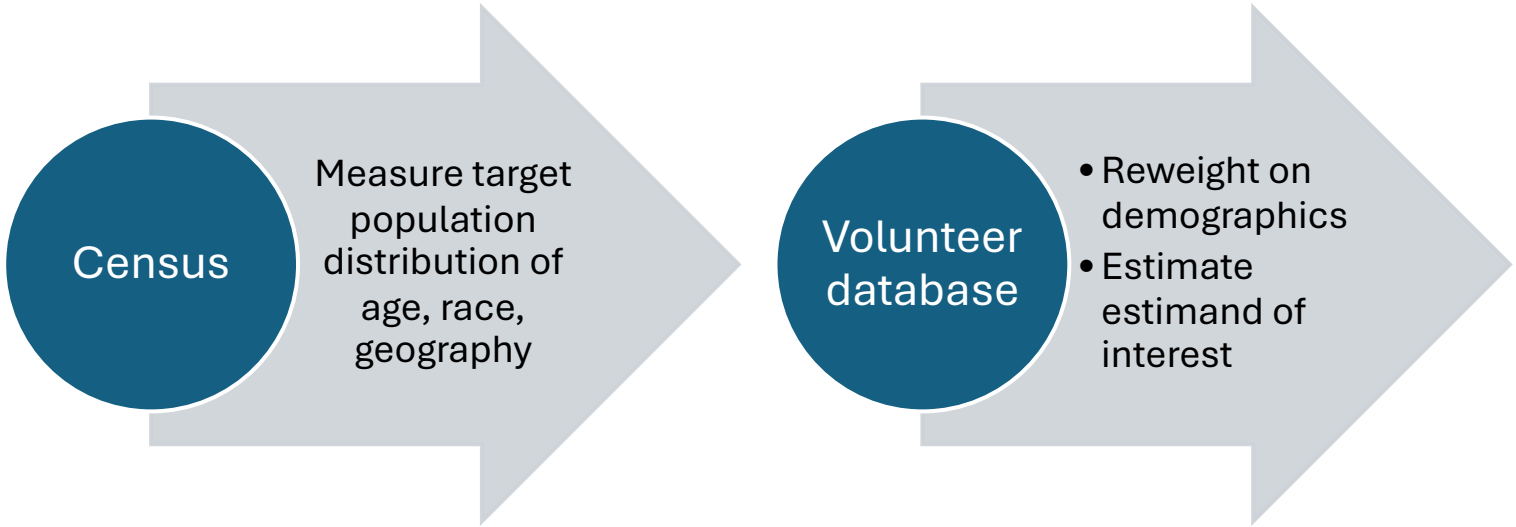
If this were stratified random sampling…

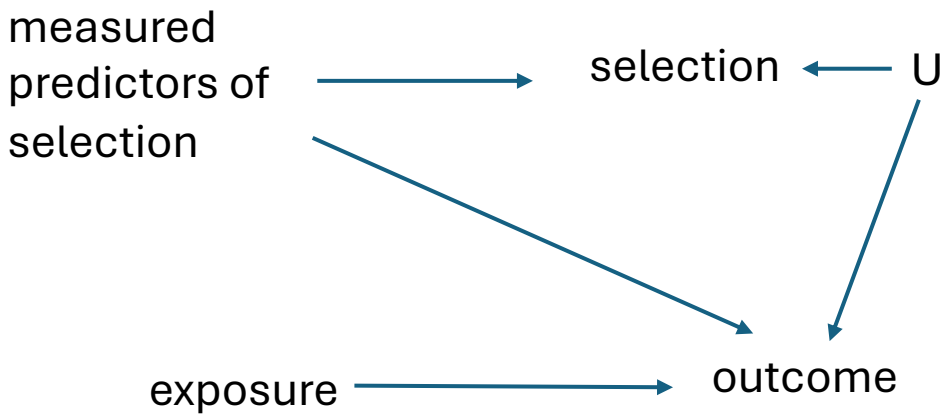
Downweight oversampled and upweight undersampled strata based on known probability of sampling within strata





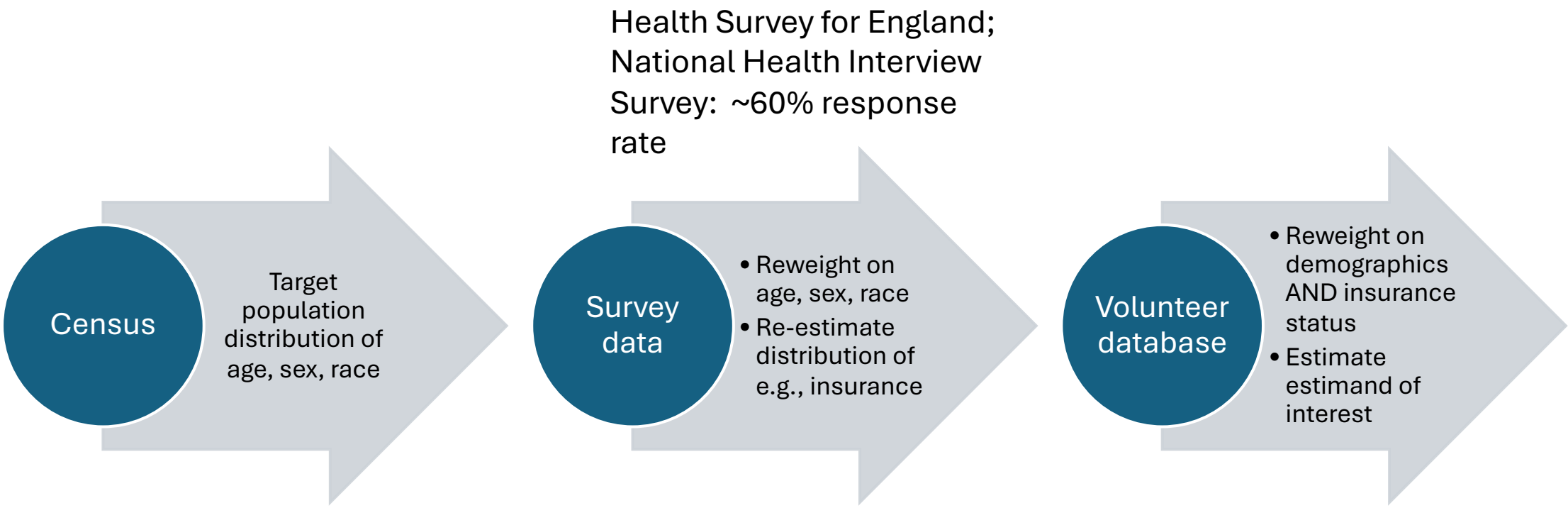
Applying the same principles to large-scale volunteer databases





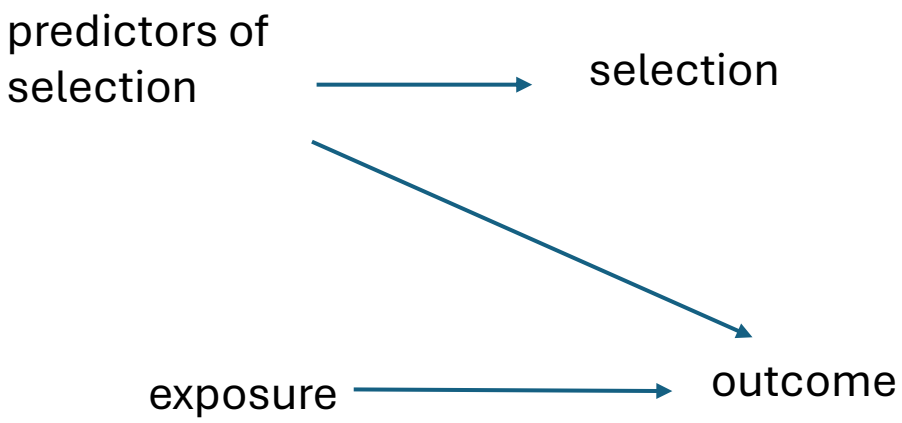
Problem: unmeasured predictors of selection

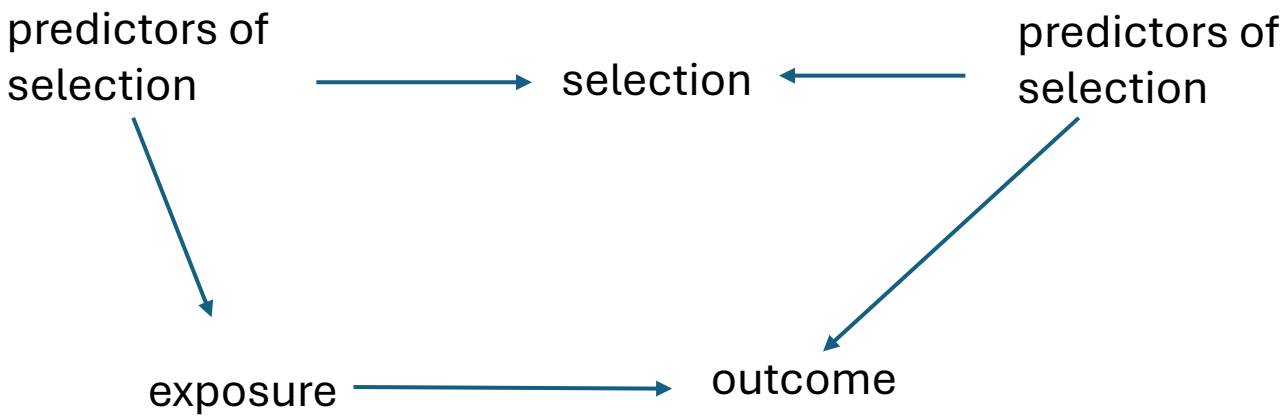
What if we could get data on some of those predictors of selection from a survey?



Problem: nationally representative surveys also suffer from selection problems!

Overarching problem? Not every question needs the same adjustment





Considerations for reweighting

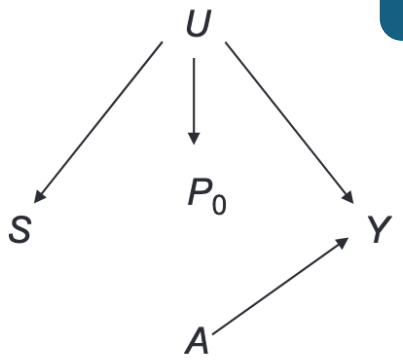
- Census/vital statistics are truly representative (mostly)
  - But lacking the rich data of cohort studies
- “Nationally representative” studies rely on their own reweighting due to non-response
  - There are likely unmeasured predictors of participation
- Think about the specific question/DAG
  - Also about your target population!
- Need positive probability of selection within all strata
- Measurement error

Avoid "type 3 selection bias" by striving for inclusiveness

Yikes! Distributions of variables are not the same – All of Us and similar studies are not representative

That's ok! We can reweight analyses to match the US (or other) population

But wait! We need to take the DAG into consideration -- the necessary weighting factors are not the same for every causal effect and may not even be measured



Vol. 192, No. 3

Practice of Epidemiology

Proxy Variables and the Generalizability of Study Results

Anton Nilsson*, Jonas Björk, and Carl Bonander

However! With comprehensive data, we may have measured enough proxies of these predictors of selection to recover causal effects – or close enough

Thanks!

l.smith@northeastern.edu