Methods for Selection Bias in the UK Biobank

Valerie Bradley

University of Oxford, Department of Statistics

15 October 2019

1. The UK Biobank
2. Why is unrepresentativeness a problem?
3. Can we recover from selection bias?
4. Methods for recovering from selection bias
5. Application to the UK Biobank

The UK Biobank

Overview

UK Biobank

- Largest ever prospective health study
- Includes genetic sequencing, blood tests, physical exams, health history questionnaire
- 500,000+ participants aged 40-70 when recruited between 2006 and 2010
- Not representative of the UK general population (Fry et al. 2017)

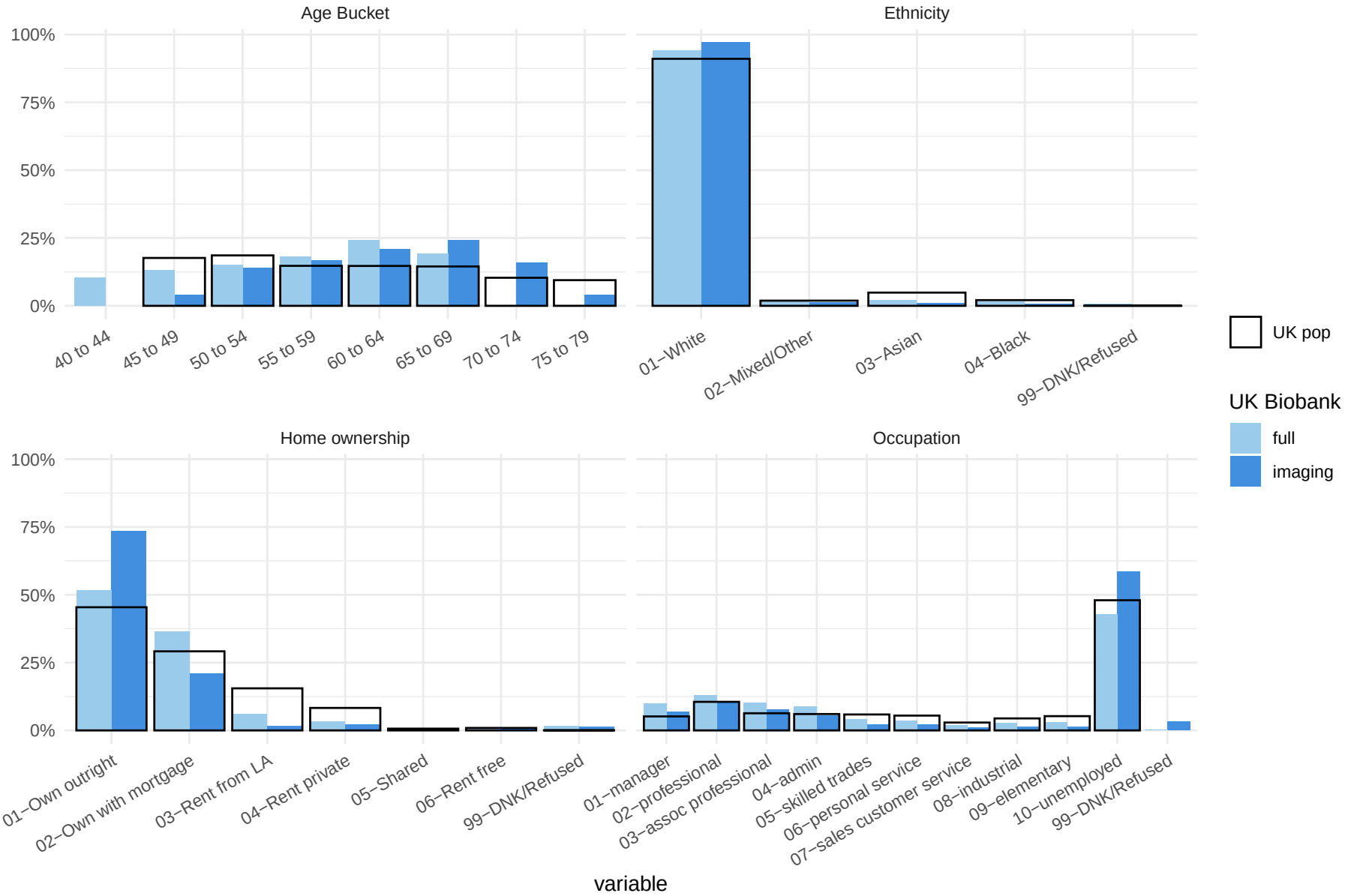
Goal: Study exposures and outcomes that affect aging populations

UK Biobank imaging cohort

- Subset of UKB participants recruited to undergo additional imaging exams
- 21,407 complete, valid T1 structural MRIs

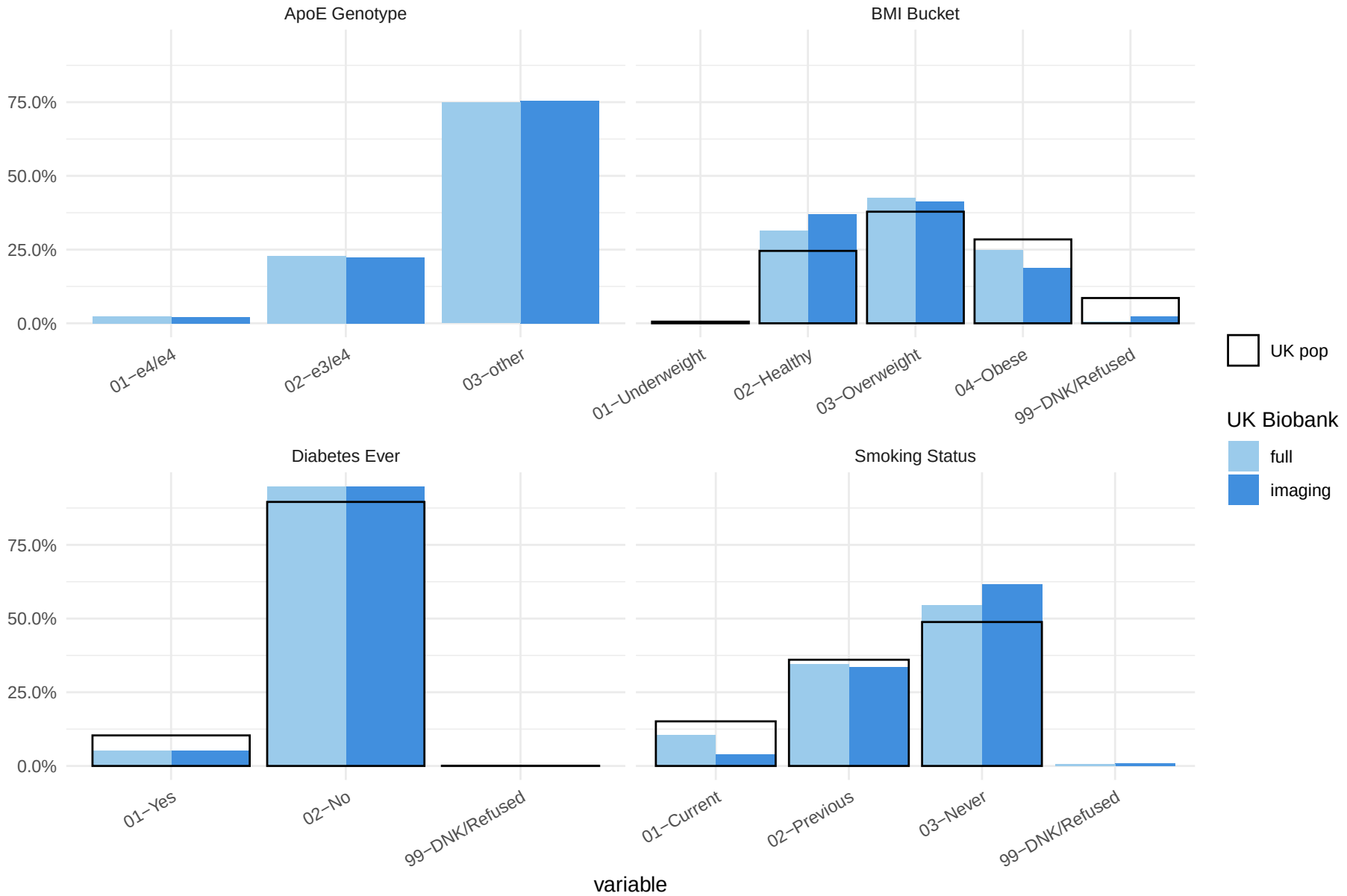
Unrepresentativeness of the UK Biobank

Demographic covariates



Unrepresentativeness of the UK Biobank

Health covariates

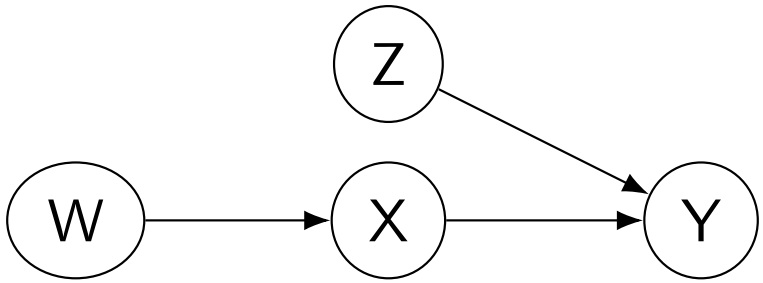


Why is unrepresentativeness a problem?

Quick note on structural causal models (SCMs)

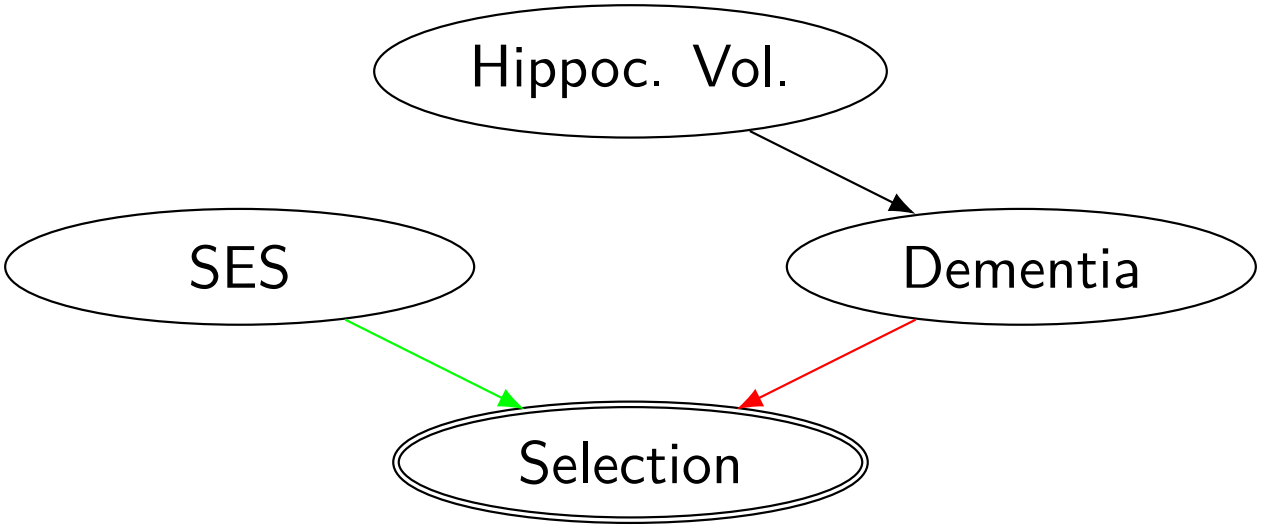
SCMs use directed acyclic graphs (DAGs) as "a mathematical language for integrating statistical and subject-matter information," specifically information about dependence structures (Pearl 1995)

- **nodes** represent *physical mechanisms*
- **edges** represent *direct causal pathways*



Note: we will focus on discrete mechanisms (e.g. $X \in \{0, 1\}$), but this can be generalized to the continuous case

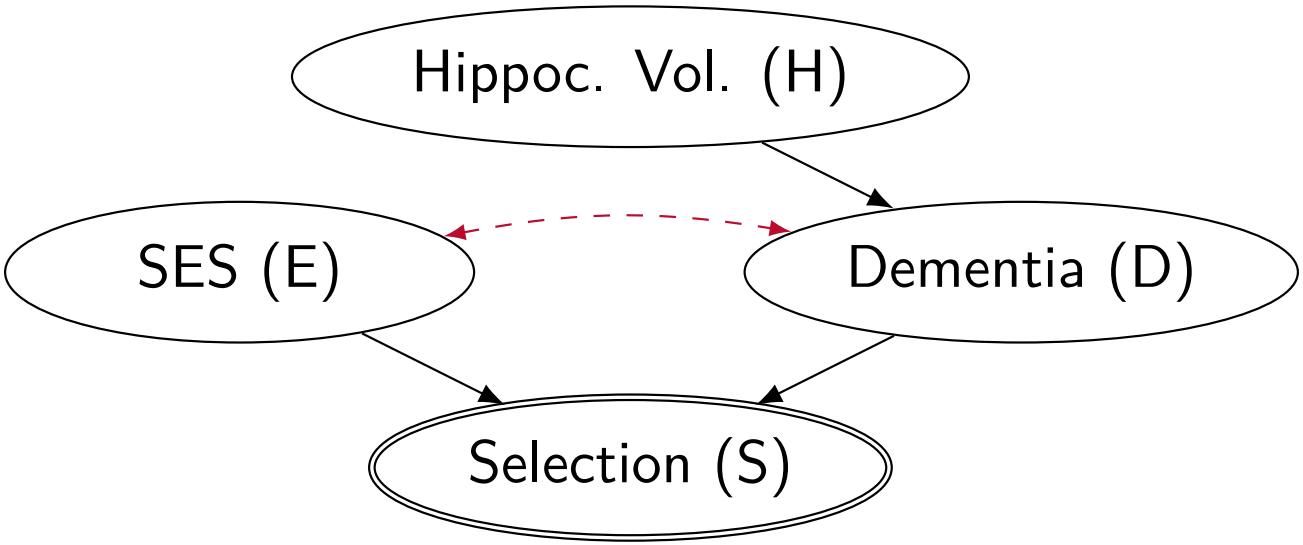
SCM Example



- SES and participation in UKB are positively correlated: Lower SES may mean it's harder to take time off of work, find childcare, etc.
- Dementia and participation are negatively correlated: Individuals with dementia (even early signs) may find it harder to make it to an appointment
- SES and hippocampal volume are independent (no edge)

What can we infer about someone with low SES who has participated in the UKB?

Collider bias



Collider bias: "When two variables independently influence a third variable, and that third variable is conditioned upon" (Munafò et al. 2018)

Opens a *backdoor path* (spurious association) between SES (E) and hippocampal volume (H).

- Want to know P(H|E)
- But only observe P($H|E, S = 1$)

Collider bias example

Day et al. (2016): GWAS using 142,630 observations from the UKB, induced strong collider bias

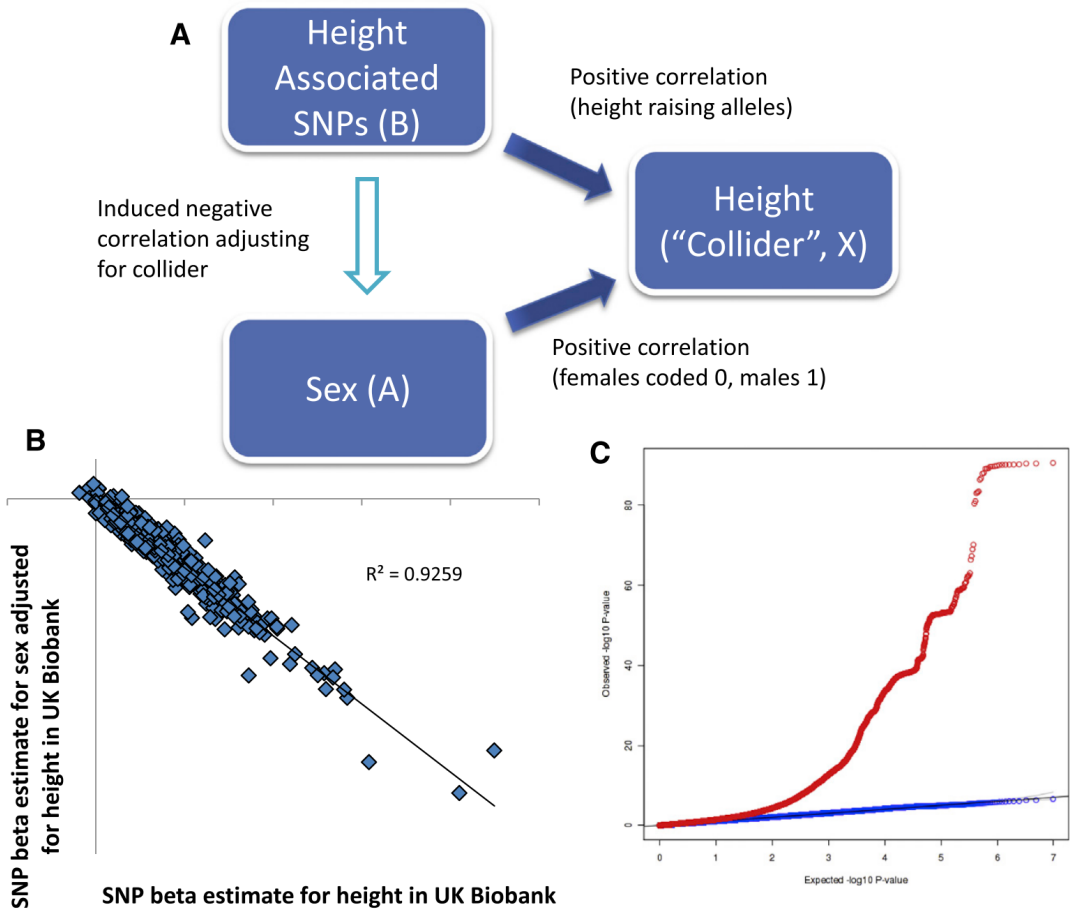


Figure 1. Induced Collider Bias between Genetic Variants, Height, and Sex
(A) Schematic diagram of the scenario in which collider bias can occur between genetic variants, height, and sex.
(B) Spurious autosomal SNP-effect estimates for sex, created by adjusting for height as a covariate, are almost perfectly correlated with SNP-effect estimates for height. In this scenario of collider bias, adjustment for the collider height creates biologically implausible sex associations for the 694 previously identified genome-wide significant autosomal SNPs for height.
(C) A quantile-quantile plot of genome-wide autosomal test statistics for sex ~ SNP (shown in blue) and sex ~ SNP + height (shown in red).

Can we recover from selection bias?

Formal conditions for recovery

Consider the *association* between X and Y, or $\mathrm{P}(Y = y|X = x)$

We can recover from selection bias **if and only if** $\mathrm{P}(y|x)$ can be written in terms of the quantities observed under selection, usually relying on a set of *auxilary variables* $\mathbf{Z}$

- if $Y \perp\!\!\!\perp S|\mathbf{Z}, X$, then we can recover the association with

$$\mathrm{P}(y|x) = \sum_{\mathbf{z}} \mathrm{P}(y|x, \mathbf{z}, S = 1)\mathrm{P}(\mathbf{z}|x)$$

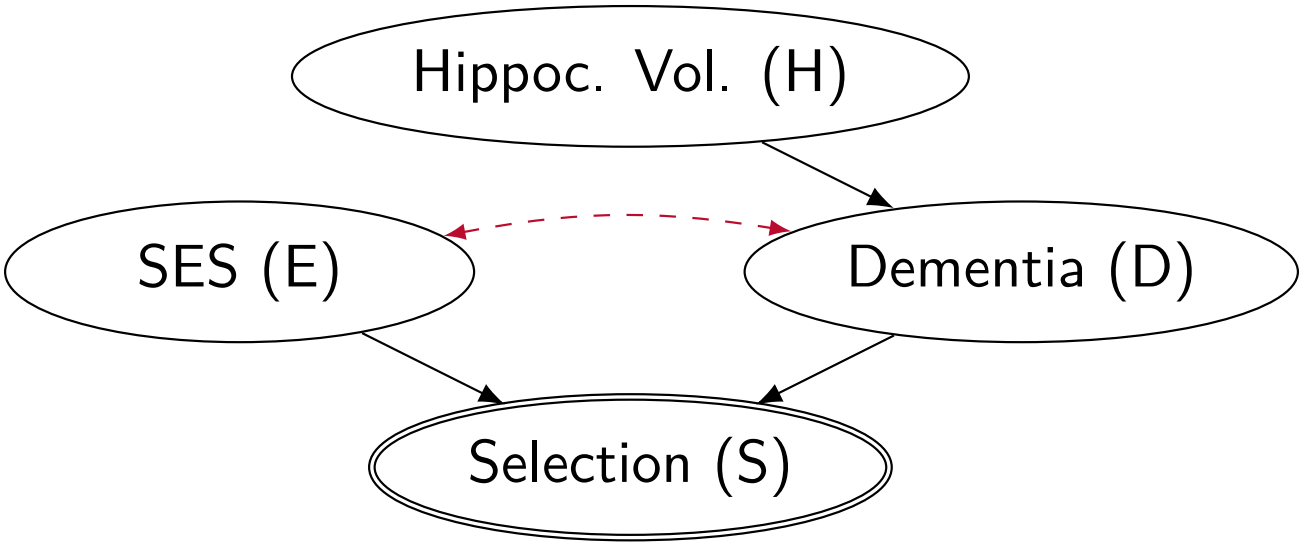
- If $\{Y \cup X\} \perp\!\!\!\perp S|\mathbf{Z}$, then we can recover the association with

$$\mathrm{P}(y|x) = \sum_{\mathbf{z}} \mathrm{P}(y|x, \mathbf{z}, S = 1)\mathrm{P}(\mathbf{z})$$

The key is that conditioning on $\mathbf{Z}$ (and maybe X) makes the outcome of interest Y *conditionally independent of selection* S

Example

Back to our example. . .



True that $H \perp\!\!\!\perp S|D, E$, but not true that $\{H \cup E\} \perp\!\!\!\perp S|D$

Therefore, we CAN recover $\mathrm{P}(H|E)$ as

$$\mathrm{P}(H|E) = \sum_{D} \mathrm{P}(H|E, D, S = 1)P(D|E)$$

only if we observe $\mathrm{P}(D|E)$ (or have an unbiased estimate)

Problems

1. Must assume we have correctly represented the dependence structures (and have correctly identified all necessary elements of $\mathbf{Z}$). This is **hard** to do in practice.
2. Don't always observe $\mathrm{P}(\mathbf{Z}|X)$ or $\mathrm{P}(\mathbf{Z})$ in full. *Can we estimate it?*
3. Can only condition on a limited number of discrete variables $\mathbf{Z}$ before this breaks down. *What happens when an element of* $\mathbf{Z}$ *is continuous?*

Methods for recovering from selection bias

Two main methods for recovery

Classes of methods for estimating effects, associations, prevalence in the presence of selection bias:

- **Inverse probability weighting (IPW)**: weight each observed unit by the inverse of their probability of selection; intuitively, creates a "pseudo-population" in which selection bias does not exist (Hernán, Hernández-Díaz, and Robins 2004)
  - con: not always clear how to incorporate into estimators
  - con: hard to correctly estimate standard errors of weighted estimators
- **Regression methods**: directly model the outcome of interest, accounting for confounders/auxiliary variables
  - con: have to estimate separate regression model for each outcome

Inverse probability weighting (IPW)

Can be derived exactly from (causal) recovery conditions (Correa, Tian, and Bareinboim 2018):

$$\mathrm{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} \mathrm{P}(\mathbf{y}|\mathbf{x},\mathbf{z},S=1)\mathrm{P}(\mathbf{z}\setminus\mathbf{z}^{\mathbf{T}}|\mathbf{z}^{\mathbf{T}},S=1)\mathrm{P}(\mathbf{z}^{\mathbf{T}}) \tag{1}$$

$$= \sum_{\mathbf{Z}} \frac{\mathrm{P}(\mathbf{y},\mathbf{x},\mathbf{z}|S=1)}{\mathrm{P}(\mathbf{x}|\mathbf{z},S=1)} \frac{\mathrm{P}(S=1)}{\mathrm{P}(S=1|\mathbf{z}^{\mathbf{T}})} \tag{2}$$

- $\mathrm{P}(\mathbf{y},\mathbf{x},\mathbf{z}|S=1)$, joint distribution of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$ under selection bias
- $\mathrm{P}(\mathbf{x}|\mathbf{z},S=1)$, the probability of treatment given covariates in the selection-biased sample data, related to the *propensity score* $\mathrm{P}(X|\mathbf{Z})$
- $\mathrm{P}(S=1)/\mathrm{P}(S=1|\mathbf{z}^{\mathbf{T}})$, the *inverse probability-of-selection weight*

Inverse probability weighting (IPW)

Say we observe outcomes Y, exposures X, and auxiliary variables Z for $i = 1, \ldots, n$ individuals. Can estimate $\mathrm{P}(Y|X)$ with

$$\hat{\mu} = E(Y|X = x) = \frac{1}{n} \sum_{i=1}^{n} w_i^s w_i^c I_{(X_i = x)} Y_i$$

- w_i^s is the inverse probability of selection weight, given $\mathbf{z}$

$$w_i^s = \hat{\mathrm{P}}(S = 1)/\hat{\mathrm{P}}(S = 1|\mathbf{Z}_i)$$

- w_i^c is the probability of treatment under selection

$$w_i^c = 1/\hat{\mathrm{P}}(X_i|\mathbf{Z}_i, S = 1)$$

- $I_{(X_i = x)}$ is an indicator function for which exposure a unit recieved

But how do we estimate the weights?

Methods: summary

Classic weighting methods:

1. *Post-stratification*: adjust to the joint distribution of $\mathbf{Z}$
2. *Raking*: iteratively adjust the marginal distributions of elements in $\mathbf{Z}$
3. *Calibration*: raking, but with continuous variables as well as discrete variables

Less-common methods:

4. *Logit*: estimate the probability of selection directly
5. *LASSO*: use a LASSO to select variables and interactions for raking

New method:

6. *BART + raking*: use a Bayesian Additive Regression Tree (BART) to estimate the probability of selection, then rake such that key marginal distributions match those of the population

Methods: Post-stratification

Adjust to the *joint* distribution of **Z**. This is exactly the definition for recovery (i.e. a sum over all combinations of levels of **Z**)

1. Define strata based on the full joint distribution of **Z**
2. Calculate the probability of selection for each stratum
3. Apply stratum-level estimates to individuals

Example:

sex	**age**	**N (sample)**	**N (pop)**	$\hat{\mathrm{P}}(S=1\|\mathbf{Z})$	w^s
Male	under 50	35	320	$\frac{35}{320}=0.109$	$\frac{0.1}{0.109}=0.917$
Male	50 plus	11	133	0.083	1.20
Female	under 50	41	355	0.115	0.870
Female	50 plus	13	192	0.068	1.47

where $\mathrm{P}(S=1)=0.1$

Then, all men under 50 in the study are given $w^s=0.917$

Methods: Post-stratification

Pros:

- quick, closed-form solution
- weighted joint distribution of $\mathbf{Z}$ exactly matches that of the population

Cons:

- $\mathbf{Z}$ must be discrete
- can only consider a limited number of $\mathbf{Z}$ before the strata get too small

Methods: Raking

Iterative proportional fitting; iteratively adjust the *marginal* distributions of auxilary variables $\mathbf{Z}$

1. Post-stratify to the population sex distribution
2. Post-stratify the *weighted* sample to the population age distribution and update the weights
3. Post-stratify the *new weighted* sample to the population sex distribution and update weights

...

Stop when weights stabilize (according to a tolerance threshold ϵ)

Methods: Raking

Pros:

- Weights are more stable, less extreme than post-stratification
- Can consider a large set of variables **Z**

Cons:

- Iterative (may never converge)
- Not considering interactions between variables

(Basically the opposite of post-stratification)

Methods: Calibration

A generalization of raking that allows for continuous **Z**

Instead of iterating over marginal distributions of elements in **Z** (i.e. P(sex) and P(age)), we iterate over the totals of each level of each element in **Z** (i.e. female, male, under 50, over 50).

With this formlation, we can also enforce constraints (weight) on **continuous** variables. * e.g. we constrain the mean age of the sample

Con: can be very finicky, even less likely to converge than raking

Methods: Directly estimate $\hat{\mathrm{P}}(S = 1|\mathbf{Z})$ with regression

$$\hat{\mathrm{P}}(S_i = 1|\mathbf{Z}_i) = \mathrm{logit}^{-1}(\beta \mathbf{Z}_i)$$

Cons:

- Weighted distribution of $\mathbf{Z}$ will almost certainly not match population distributions (making results much less interpretable)
- Requires individual-level population data

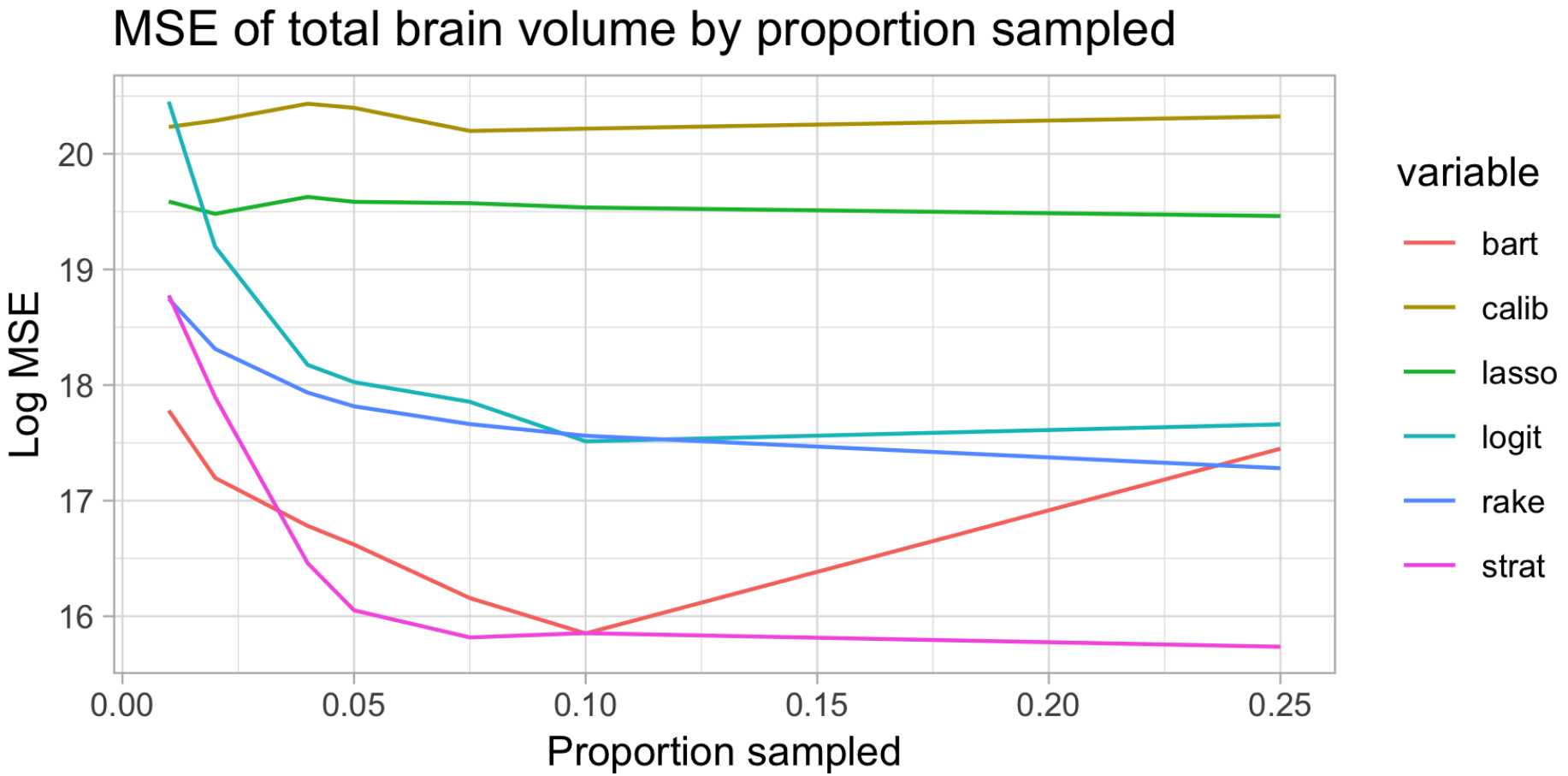
Methods: Raking with LASSO variable selection

Problem: Both raking and post-stratification can fail when $\mathbf{Z}$ is too large

Solution: Select significant subsets of $\mathbf{Z}$ using LASSO

Application to the UK Biobank

Simulation Results

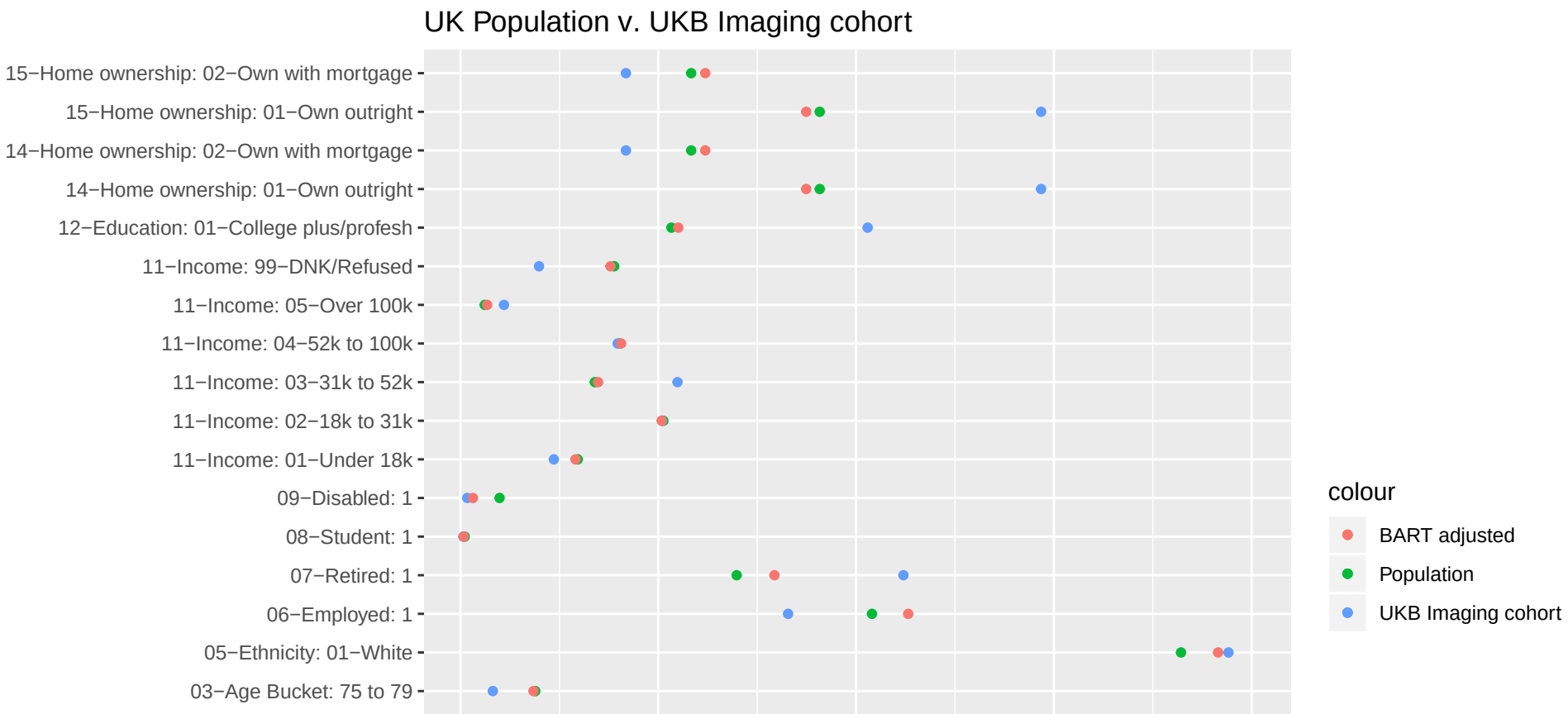


Application to UK Biobank

```
## Warning: Removed 1 rows containing missing values (geom_poi

## Warning: Removed 1 rows containing missing values (geom_poi

## Warning: Removed 1 rows containing missing values (geom_poi
```



Summary

- The UKB and the imaging cohort are not representative of the UK general population
- This is a problem because **collider bias** can impact estimates of association
- Reviewed SCMs as a language for expressing dependence structures

References I

Correa, J D, J Tian, and E Bareinboim. 2018. "Generalized adjustment under confounding and selection biases." *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, no. June: 6335–42. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059403683%7B/&%7DpartnerID=40%7B/&%7Dmd5=d87612fd7a9f1c4cc8831887080291ab.

Day, Felix R., Po Ru Loh, Robert A. Scott, Ken K. Ong, and John R. B. Perry. 2016. "A Robust Example of Collider Bias in a Genetic Association Study." *American Journal of Human Genetics* 98 (2): 392–93. https://doi.org/10.1016/j.ajhg.2015.12.019.

Fry, Anna, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. 2017. "Comparison of Sociodemographic and Health-Related Characteristics of Uk Biobank Participants with Those of the General Population." *American Journal of Epidemiology* 186 (9): 1026–34.

References II

Hernán, Miguel A., Sonia Hernández-Díaz, and James M. Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15 (5): 615–25. https://doi.org/10.1097/01.ede.0000135174.63482.43.

Munafò, Marcus R., Kate Tilling, Amy E. Taylor, David M. Evans, and George Davey Smith. 2018. "Collider scope: When selection bias can substantially influence observed associations." *International Journal of Epidemiology* 47 (1): 226–35. https://doi.org/10.1093/ije/dyx206.

Pearl, Judea. 1995. "Causal diagrams for empirical research." *Biometrika* 82 (4): 669–710. https://doi.org/10.1093/biomet/82.4.700.