Methods for Selection Bias in the UK Biobank

Valerie Bradley

University of Oxford, Department of Statistics

15 October 2019

Unrepresentativeness in the UK Biobank

About the data

UK Biobank

- **Goal**: Study exposures and outcomes that affect aging populations
- 500,000+ participants aged 40-70 when recruited between 2006 and 2010
- *Not representative of the UK general population* (Fry et al. 2017)

UK Biobank imaging cohort

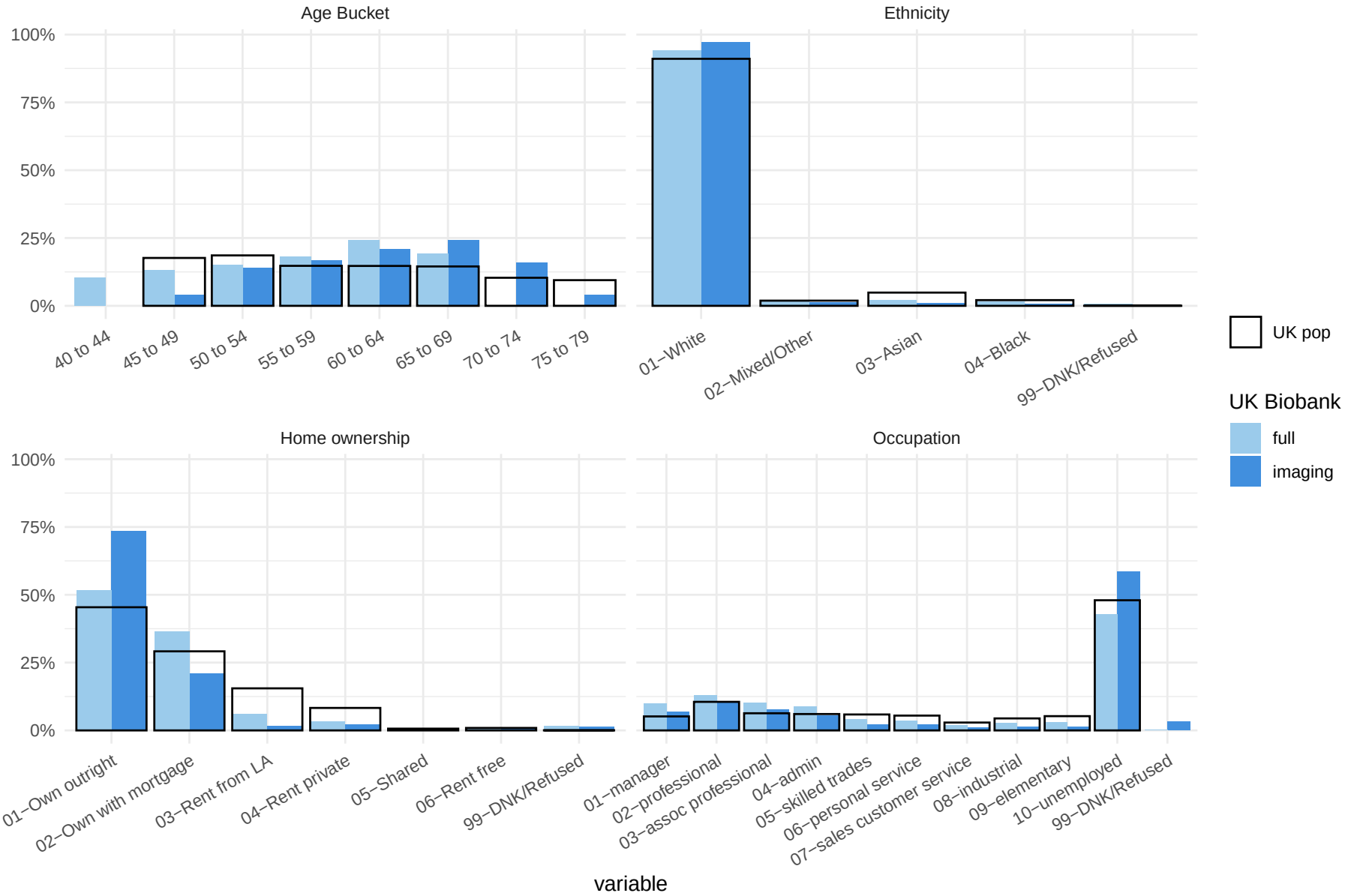
- Subset of UKB participants, undergo additional imaging exams
- 21,407 complete, valid T1 structural MRIs (4% of UKB)

2016 Health Survey for England

- Designed to estimate prevalence of health outcomes (e.g. smoking, obesity, high blood pressure)
- 2016 survey contains 10,067 respondents, 4,318 aged 45-80

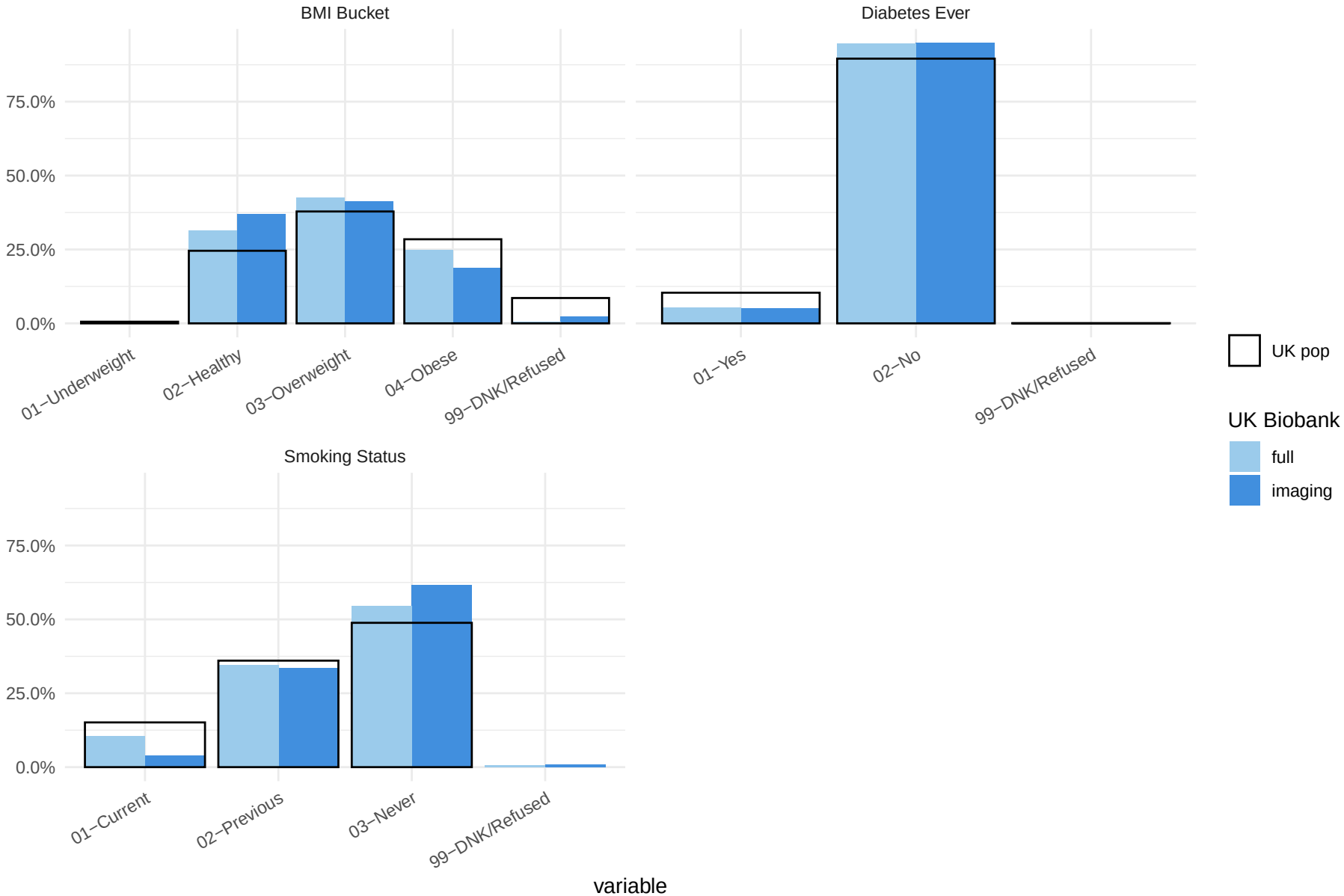
Quantifying unrepresentativeness of the UKB

Demographic covariates



Quantifying unrepresentativeness of the UKB

Health covariates

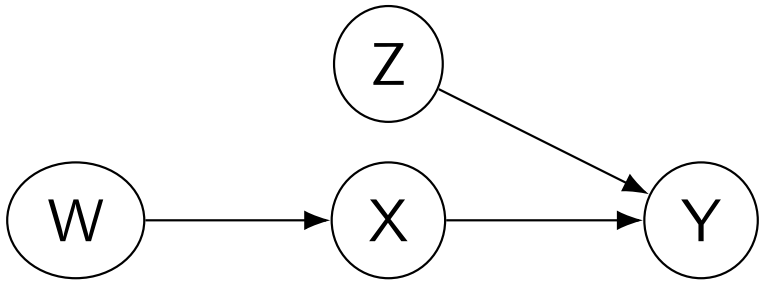


Why is unrepresentativeness a problem?

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)

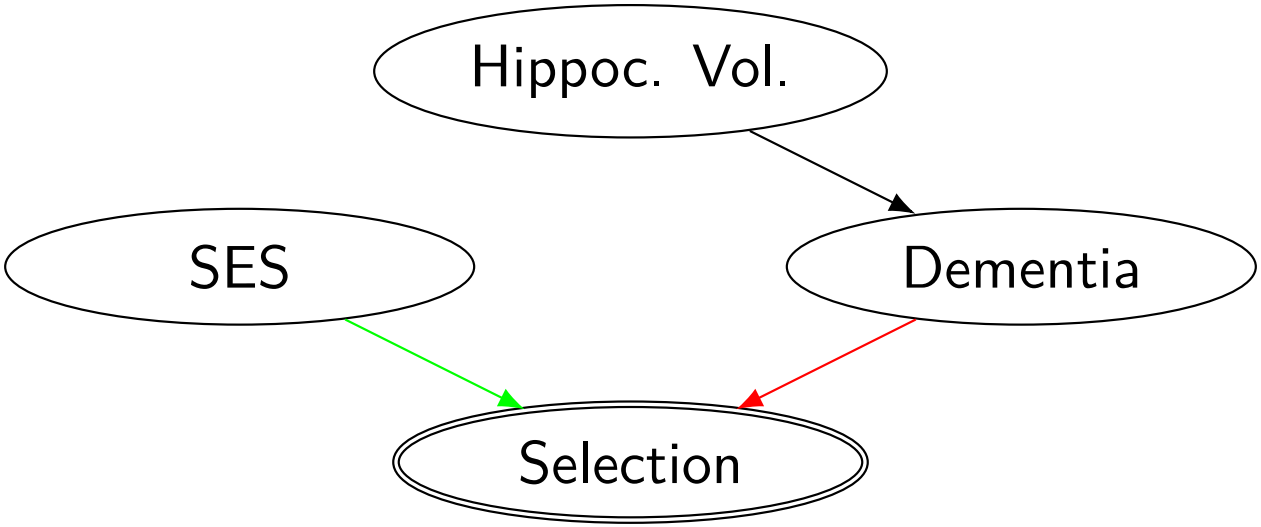
- **nodes** represent *physical mechanisms*
- **edges** represent *direct causal pathways*



Things we can say:

- there is a *direct* causal pathway from X to Y
- there is an *indirect* causal pathway from W to Y
- X *d-separates* W and Y, such that $W \perp\!\!\!\perp Y | X$

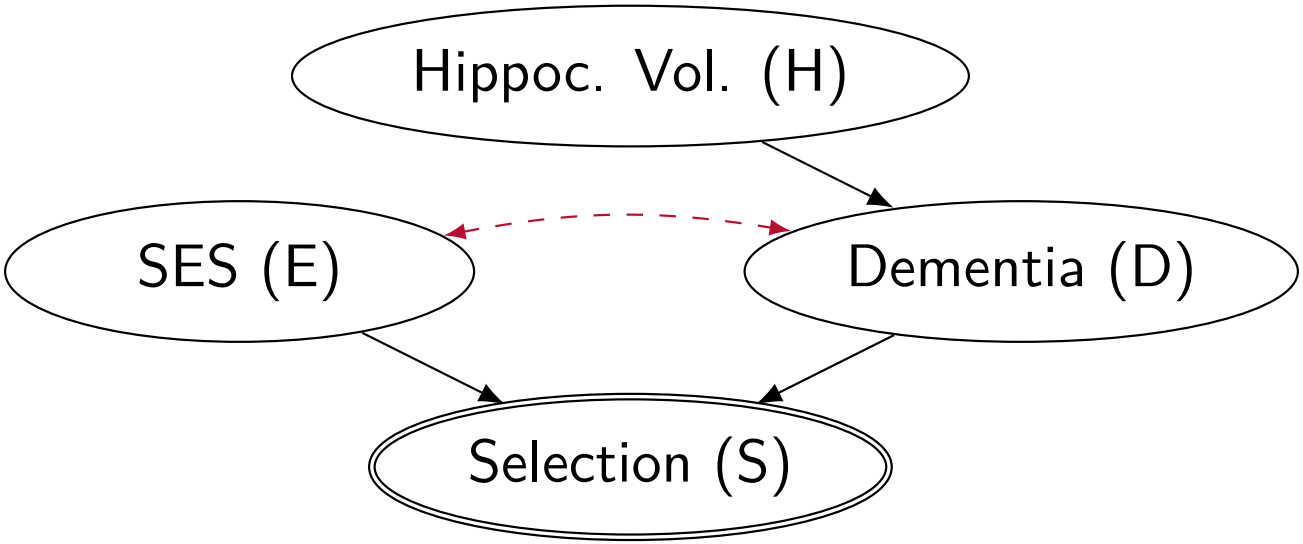
SCM Example



- SES and participation in UKB are positively correlated
- Dementia and participation are negatively correlated
- No causal relationship between SES and hippocampal volume (no edge)

Are SES and Dementia independent?

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 E and H.

- Want to know P(H|E)
- But only observe P($H|E, S = 1$)
- And, because H, E are *not* independent of S, those are not the same

Collider bias example

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

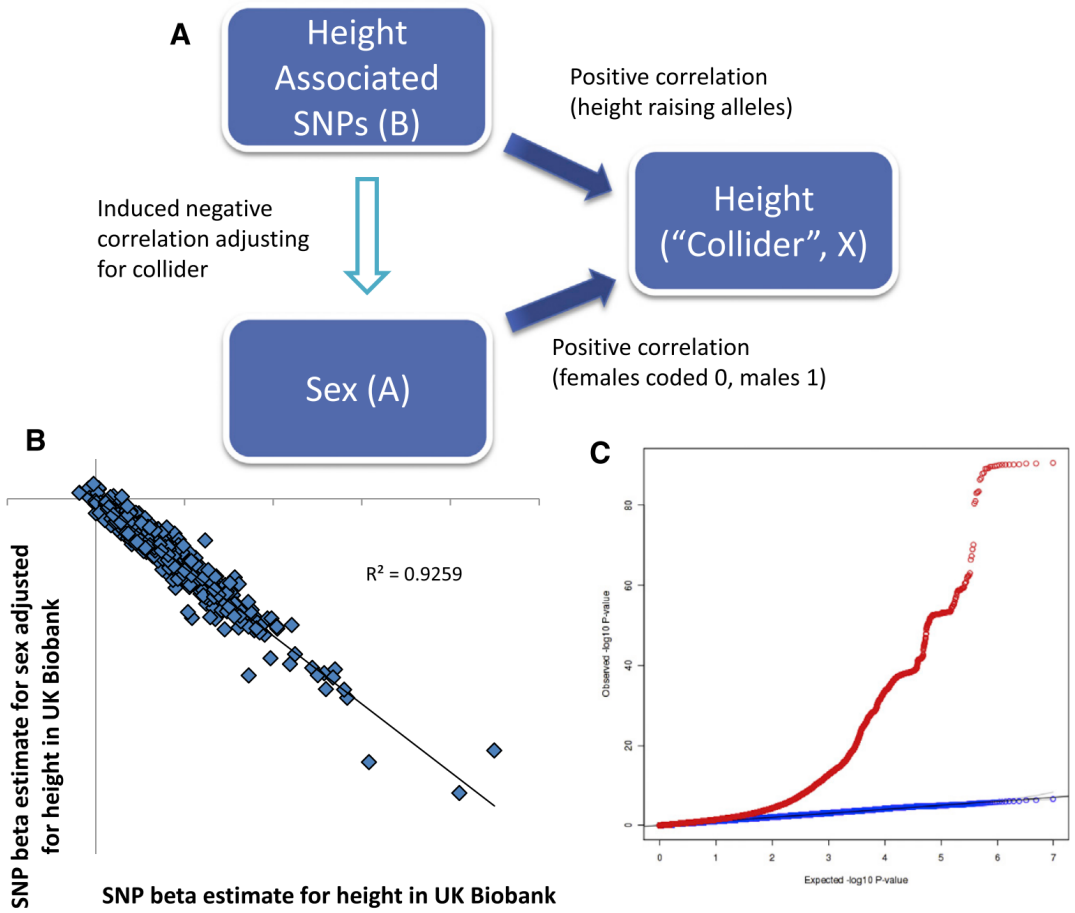


Figure 1. Induced Collider Bias between Genetic Variants, Height, and Sex
(A) Schematic diagram of the scenario in which collider bias can occur between genetic variants, height, and sex.
(B) Spurious autosomal SNP-effect estimates for sex, created by adjusting for height as a covariate, are almost perfectly 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, $\mathrm{P}(y|x)$

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; use *auxilary variables* $\mathbf{Z}$

- if $Y \perp\!\!\!\perp S|\{\mathbf{Z}, X\}$, then

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

- If $\{Y \cup X\} \perp\!\!\!\perp S|\mathbf{Z}$, then

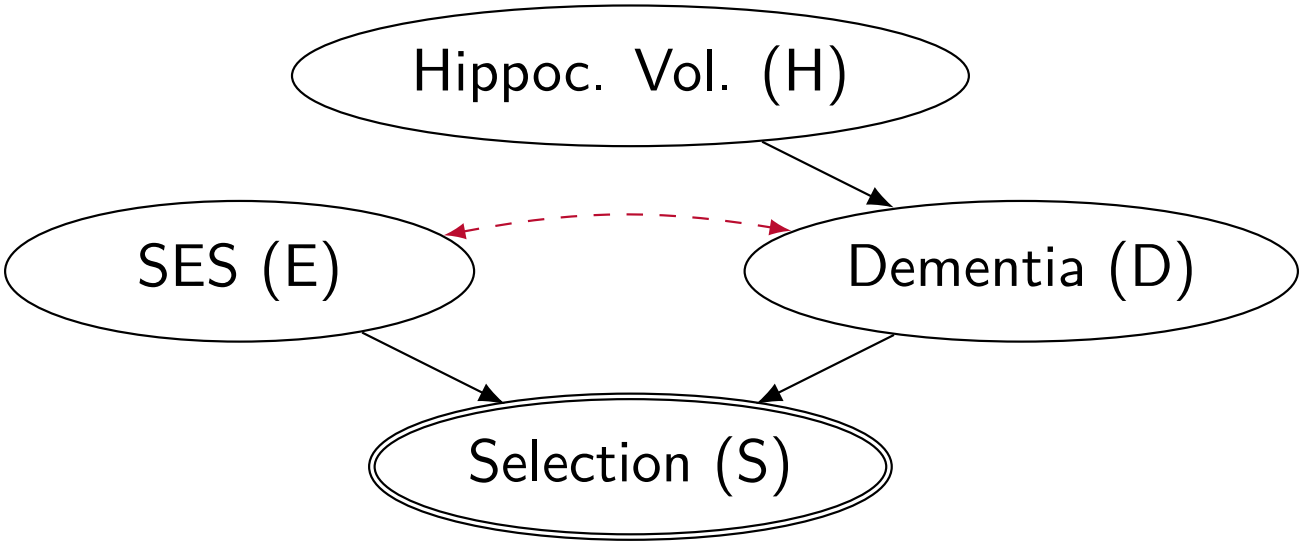
$$\mathrm{P}(y|x) = \sum_{\mathbf{z}} \mathrm{P}(y|x, \mathbf{z}, S = 1)\mathrm{P}(\mathbf{z})$$

The key: conditioning on $\mathbf{Z}$ (and maybe X) makes Y *conditionally independent of selection* S

$\mathrm{P}(\mathbf{z})$ and $\mathrm{P}(\mathbf{z}|x)$ are population distributions

Example

Back to our example. . .



- ☒ $\{H \cup E\} \perp\!\!\!\perp S|D$
- ☑ $H \perp\!\!\!\perp S|D, E$

So, can only recover $\mathrm{P}(H|E)$ as

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

as long as we observe $\mathrm{P}(D|E)$

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 the probability of selection, $w_i \propto 1/\mathrm{P}(S = 1|\mathbf{Z}_i)$
  - pro: weights are independent of Y and X
  - 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 $\mathrm{P}(Y|X)$, accounting for confounders/auxiliary variables
  - con: have to estimate separate 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*

But how should we estimate $\mathbf{P}(S=1|\mathbf{Z})$?

Estimating $P(S = 1|\mathbf{Z})$

Things to think about when comparing methods for estimating $P(S = 1|\mathbf{Z})$:

1. Weighted **marginal distribution** of $\mathbf{Z}$ should match that of the population
2. Avoid extreme weights that inflate the **variance** of estimators (have to account for the additional uncertainty of weights)
3. **Computational complexity** of the method
4. How well does the method account for interactions between elements of $\mathbf{Z}$

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

Application to the UK Biobank

Simulation overview

1. Generate a probability of missingness p_i for all 21,407 imaging subjects in the UKB
   - p_i depends on covariates that we know to be related to brain volume (mainly age) so that samples are biased
2. For sample sizes in $n_{sim} = 21,407 \times (0.01, 0.02, 0.04, 0.05, 0.075, 0.1, 0.25)$:
   1. Draw sample of size n_{sim} from imaging cohort with probability proportional to p_i
   2. Weight sample to $P(\mathbf{Z})$ defined by UKB imaging cohort using each of 6 methods
   3. Perform steps 2.1-2.2 1000 times

Simulation overview

Evaluate methods based on:

- **design effect**: measures the decrease in effective sample size (ESS) from weighting

$$\text{deff}(\mathbf{w}) = 1 + \text{Var}(\mathbf{w})$$

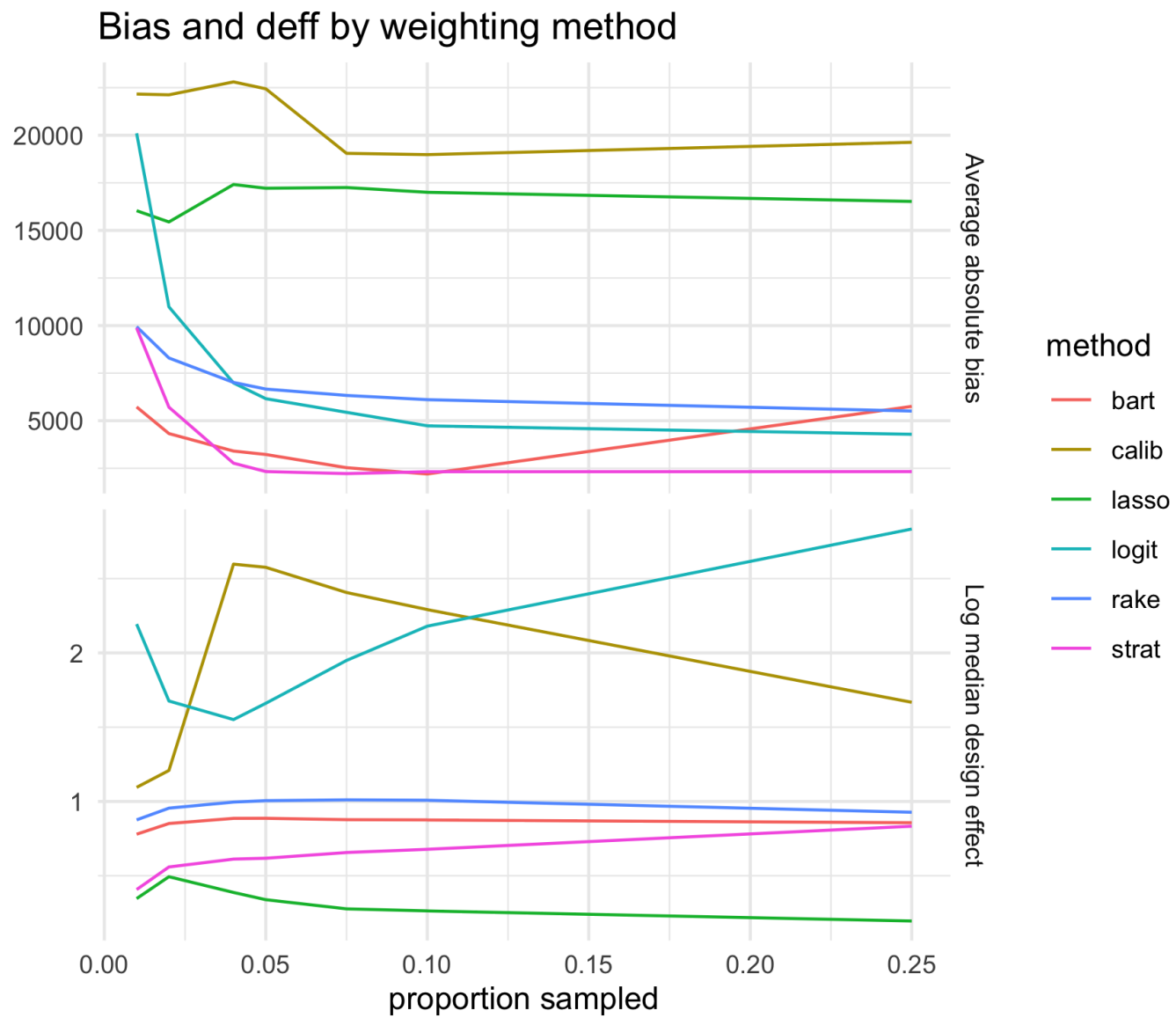
$$ESS = \frac{n}{\text{deff}(\mathbf{w})}$$

- **Absolute bias** of estimated average total brain volume

$$\text{bias} = |\bar{Y}^w - \mu|$$

$$\bar{Y}^w = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} Y_i \hat{w}_i$$

Simulation results

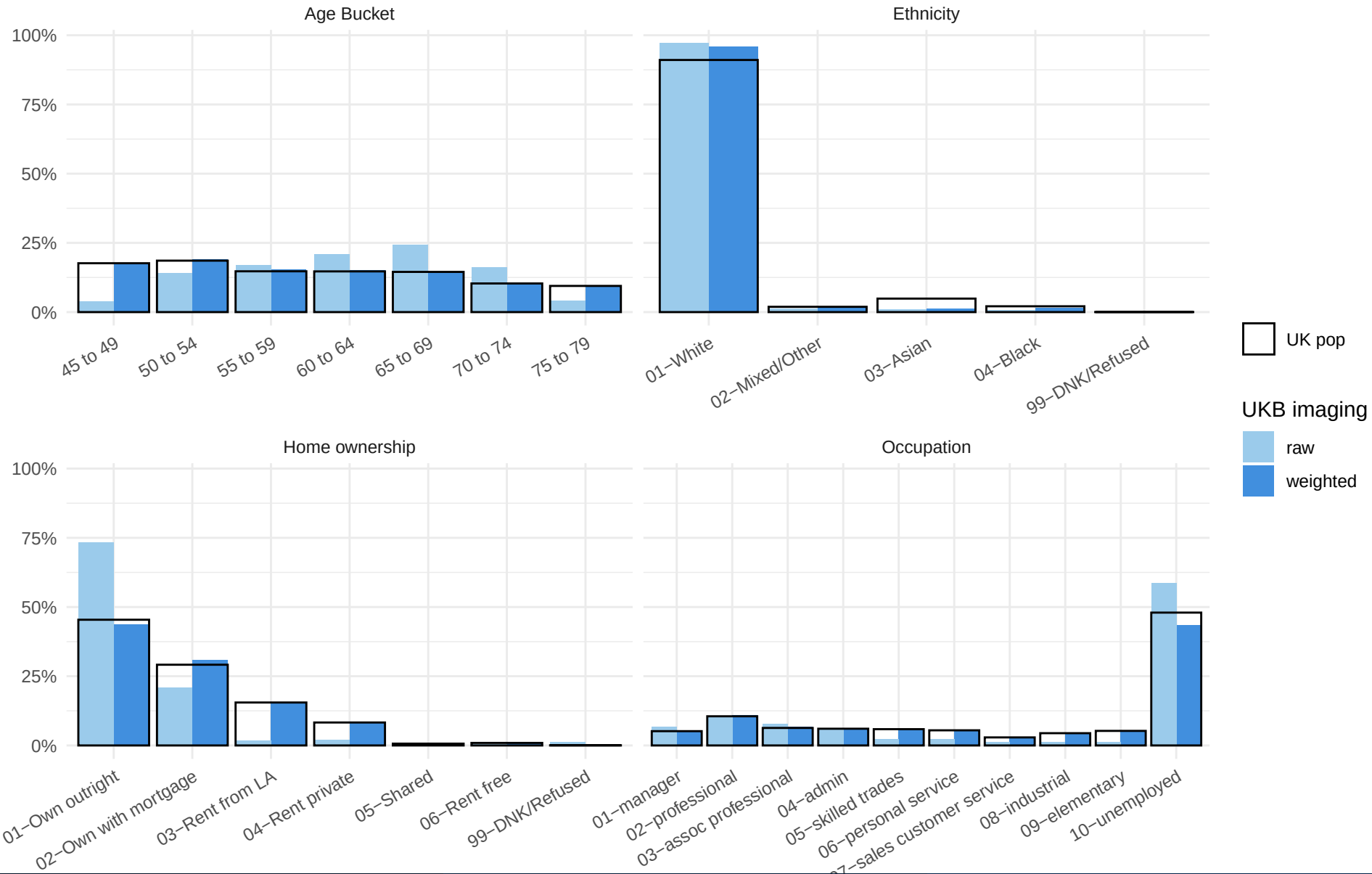


Simulation results - Remarks

- *Post-stratification* performs surprisingly well, likely due to the size of the data set
  - note that stratification variables chosen with a random forest (not the standard implementation)
- *BART + raking* outperforms other methods at the smallest sample sizes (which is probably the most realistic setting)
- *Calibration* performs very poorly - fails to correct bias and has large variance (likely from sensitivity to age distribution)
- Directly predicting $P(S = 1|\mathbf{Z})$ with *logistic regression* corrected bias well, but at a rather large cost in variance
- *LASSO* has smallest deff, but also fails to correct bias (variable selection not working well)

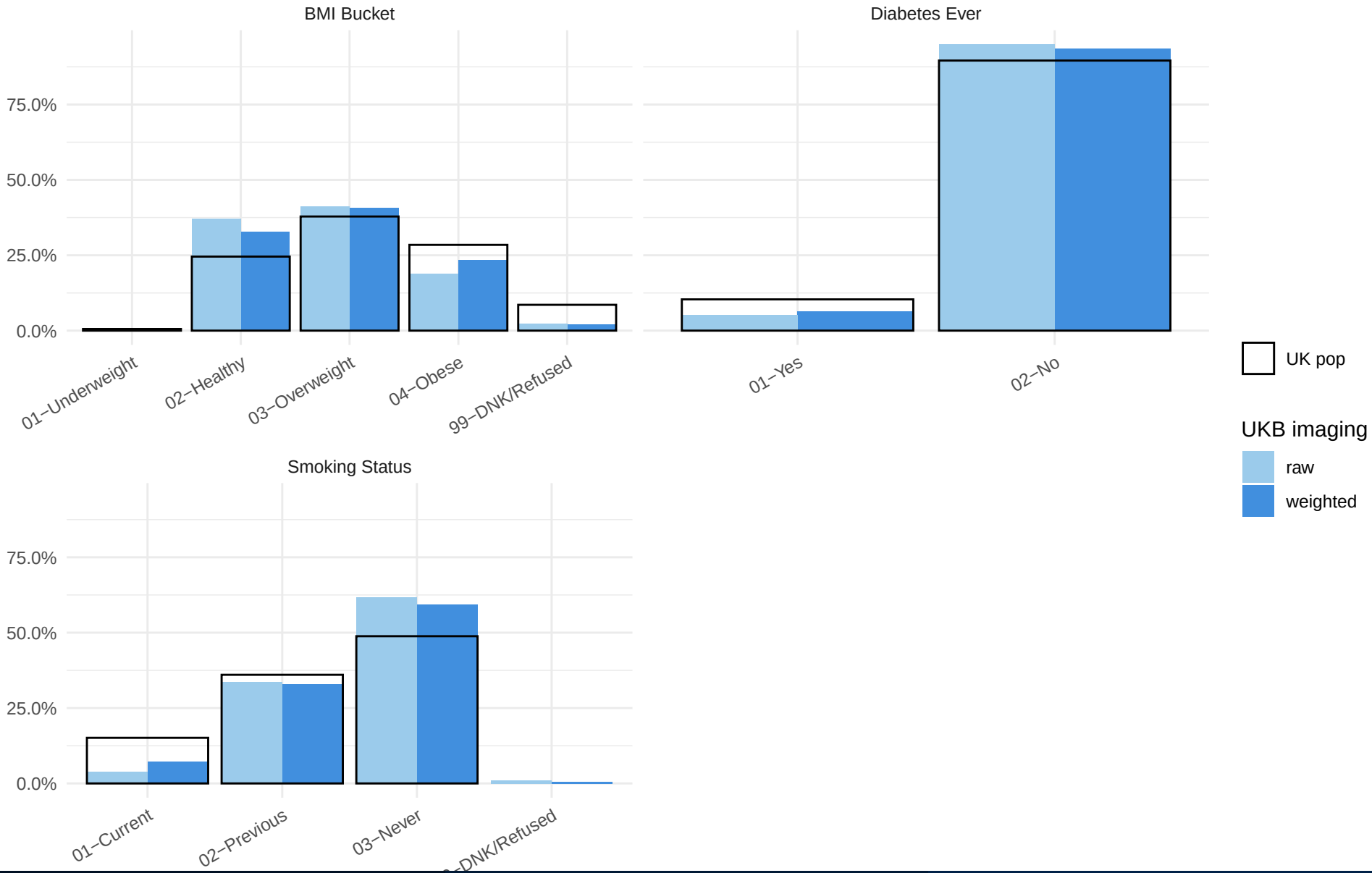
Application to UK Biobank imaging cohort

Weighted demographic distributions

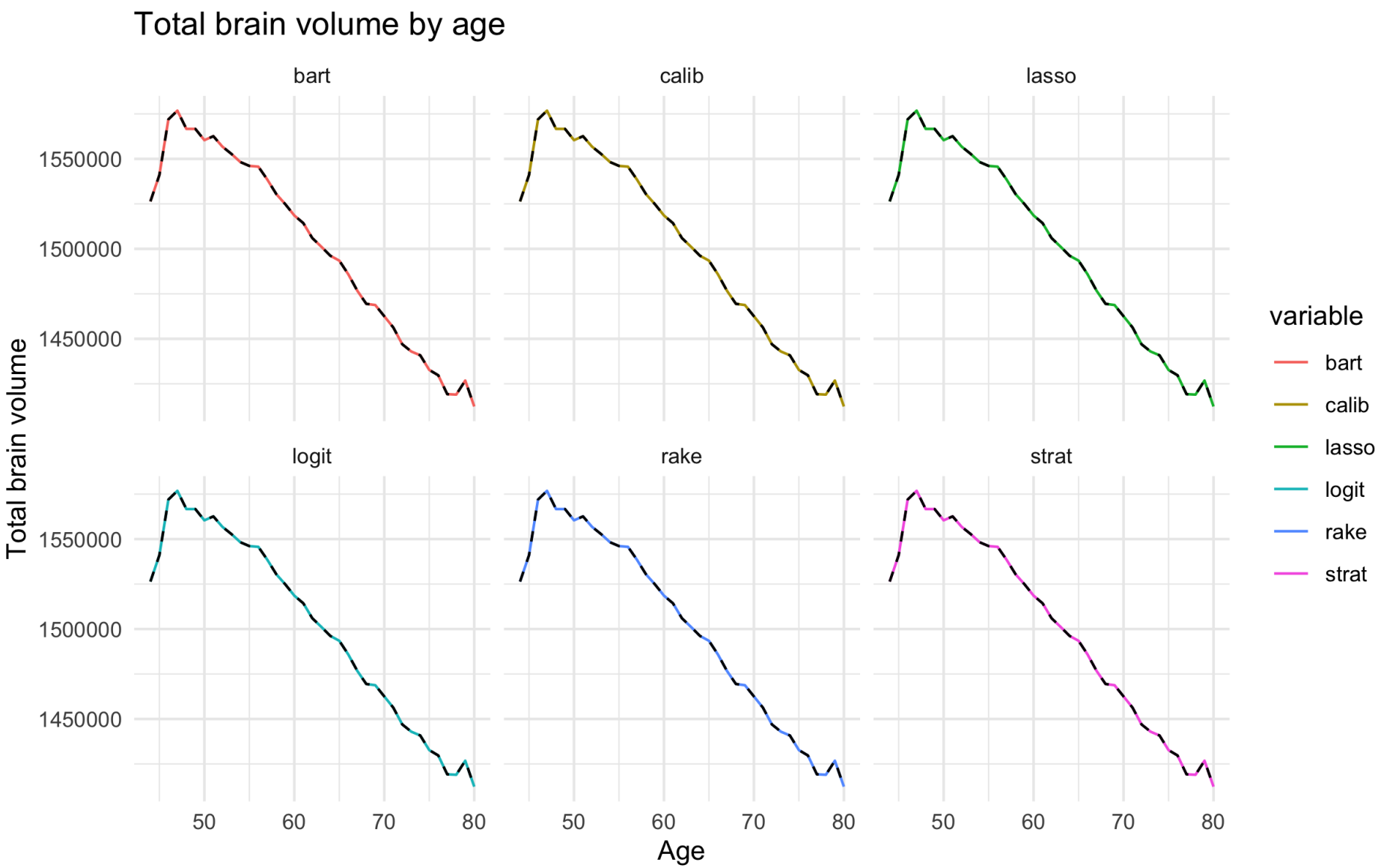


Application to UK Biobank imaging cohort

Weighted health outcome distributions



Application to the UK Biobank imaging cohort



Summary

- Be wary of collider bias when estimating associations in the UK Biobank!
- SCMs are a good framework for thinking about dependence structures, potential sources of bias, and if that bias can be corrected
- **inverse probability weighting** can be used to adjust for selection bias, many different ways to estimate weights
- Weighting the UKB imaging cohort corrects most of the bias in the demographic covariates, but only some of the bias in health outcomes

Appendix

Method 1: 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 $w^s=0.917$

Method 1: 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

Method 2: 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)

Method 2: 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)

Method 3: 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

Method 4: Directly estimate $\hat{\mathrm{P}}(S = 1|\mathbf{Z})$ with regression

We use logistic regression to estimate $\hat{\mathrm{P}}(S_i = 1|\mathbf{Z}_i)$ because selection is binary (S = 1 if observed, 0 otherwise):

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

Pros:

- Can account for a large $\mathbf{Z}$, including continous variables and interactions
- Don't need custom weighting tools, just logistic regression

Cons:

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

Method 5: 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, then rake to those marginals (Caughey and Hartman 2017)

General procedure:

1. Specify all levels of $\mathbf{Z}$ and subsets to consider (i.e. all first-order terms, and maybe two-way interactions)
2. Fit LASSO to $S_i = \text{logit}^{-1}(\beta \mathbf{Z}_i)$
3. Fit LASSO to $Y_i = f(\beta \mathbf{Z}_i)$ (f depends on likelihood of Y)
4. Rake to marginal distributions of all levels of $\mathbf{Z}$ for which the corresponding $\beta \neq 0$ in either of the LASSOs

Caution: Highly dependent on LASSO performance

Method 6: BART + raking

Intuition: Use **Bayesian additive regression tree (BART)** to estimate $P(S = 1|\mathbf{Z})$, then rake to selected $\mathbf{Z}$ so that marginal distributions match (for interpretability)

Why a BART?

- Trees are great for interactions
- Some parallels to post-stratification

(but could use any method)

Caution: Computation time much greater than other methods

References I

Caughey, Devin, and Erin Hartman. 2017. "Target Selection as Variable Selection : Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights."

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.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17375/16207.

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

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.