The effects of demographic-based selection bias on GWAS results in the UK Biobank

Sjoerd van Alten, Benjamin W. Domingue, Titus Galama, Andries T. Marees

Research Question

Does the non-random sample selection of the UK Biobank (UKB) cause selection bias in association statistics?

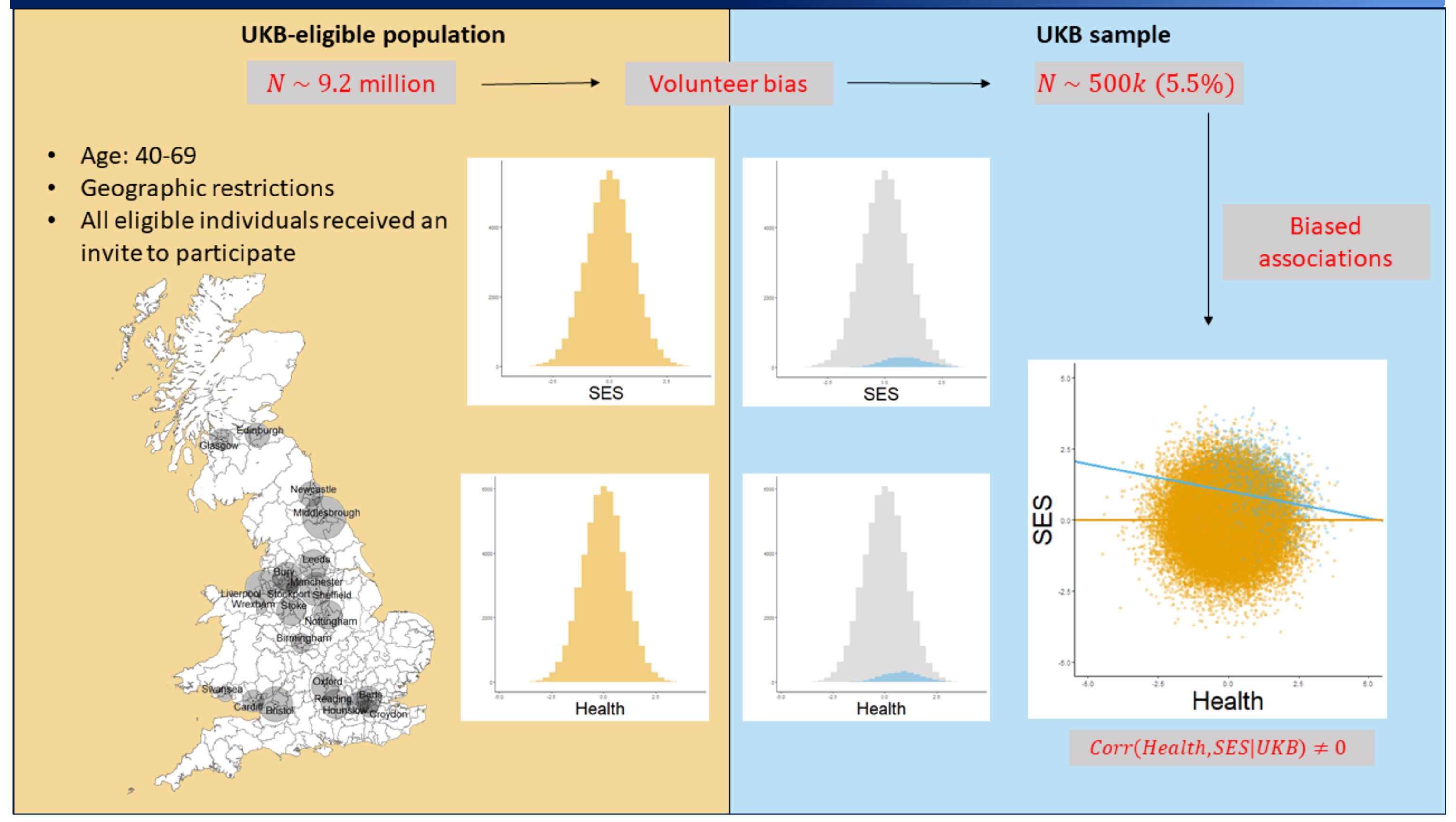
Literature review

- Large cohorts necessary for GWAS, but often non-randomly selected
- `Healthy volunteer bias' in many GWAS cohorts including UKB
- Selection bias may lead to false positive associations between genetic variants and phenotypes
  - E.g., sex shows significant autosomal heritability in the UKB, which can be attributed to selection bias

Contributions

- Demonstrate non-random selection into the UKB causes significant bias in association statistics
- Weight the UKB to make it representative of its underlying population and estimate GWAS results robust to non-random selection (education, BMI, and height: more phenotypes to be added later)
  - PGSs for education and BMI become more predictive after adjustment for volunteer bias
  - Weighted SNP associations robust to volunteer bias show stronger effect sizes for top 5,000 GWAS hits for these phenotypes

Non-random sample selection in the UKB biases association statistics (a simulated example)



UKB-eligible population

$N \sim 9.2$ million → Volunteer bias → $N \sim 500k$ (5.5%)

UKB sample

- Age: 40-69
- Geographic restrictions
- All eligible individuals received an invite to participate

Biased associations

$Corr(Health, SES|UKB) \neq 0$

Figure 1

Data

- 5% 2011 UK Census data as a population-representative reference sample
  - Create subsample representative of the UKB-eligible population:
    1. Only individuals born 1936-1970
    2. Only individuals living in grouped local authorities from which UKB sampled respondents (fig. 1)
    3. Only individuals who reported being of white ethnicity
- UKB: only keep genotyped respondents of white British ancestry whose genotyped data passes QC

Method

- Model selection into the UKB:
  - $\Pr(UKB = 1|\, Z_i') = \Phi(\alpha + Z_i'\delta + \nu_i)$
  - Z_i' includes 5-year birth cohort, sex, education, Census region, self-reported health, tenure of dwelling, employment status
- $IPW_i = \frac{\widehat{Pr}(UKB=1)}{\widehat{Pr}(UKB=1|Z_i')}$
- Use IPW_i in weighted regression of associations estimated in UKB
- Trim weights: set values in the tails equal to 1st or 99th percentile

Results (1)

- Selection into the UKB biases association statistics (fig. 2)
- Weighted regression in the UKB recovers the population-representative estimate (fig. 2)

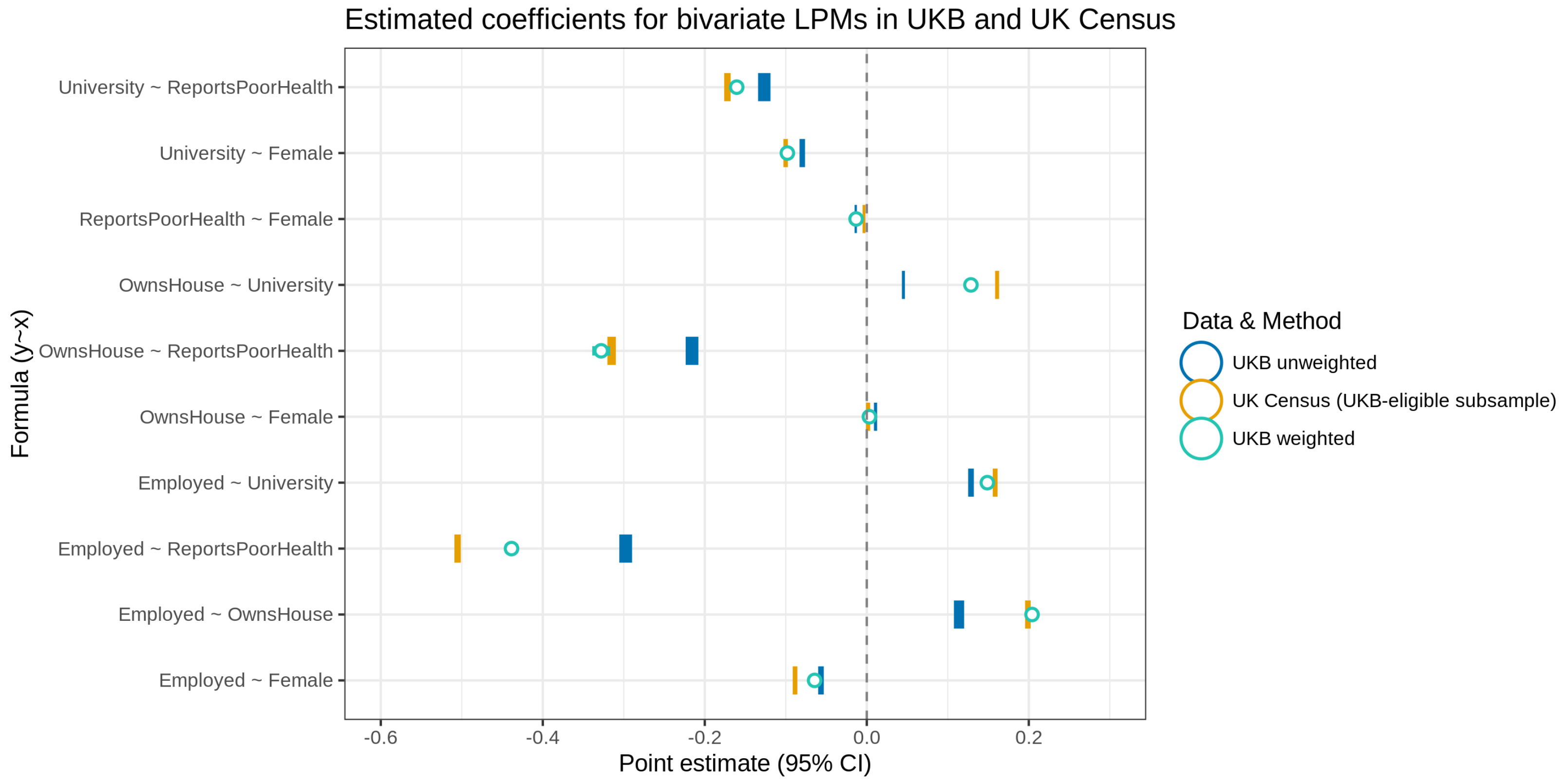


Figure 2

Results (2)

- Correcting for volunteer bias reveals larger predictiveness of PGSs for behavioral traits (EA and BMI), but not height (fig. 3)
- The PGSs are constructed using GWASs that did not include the UKB

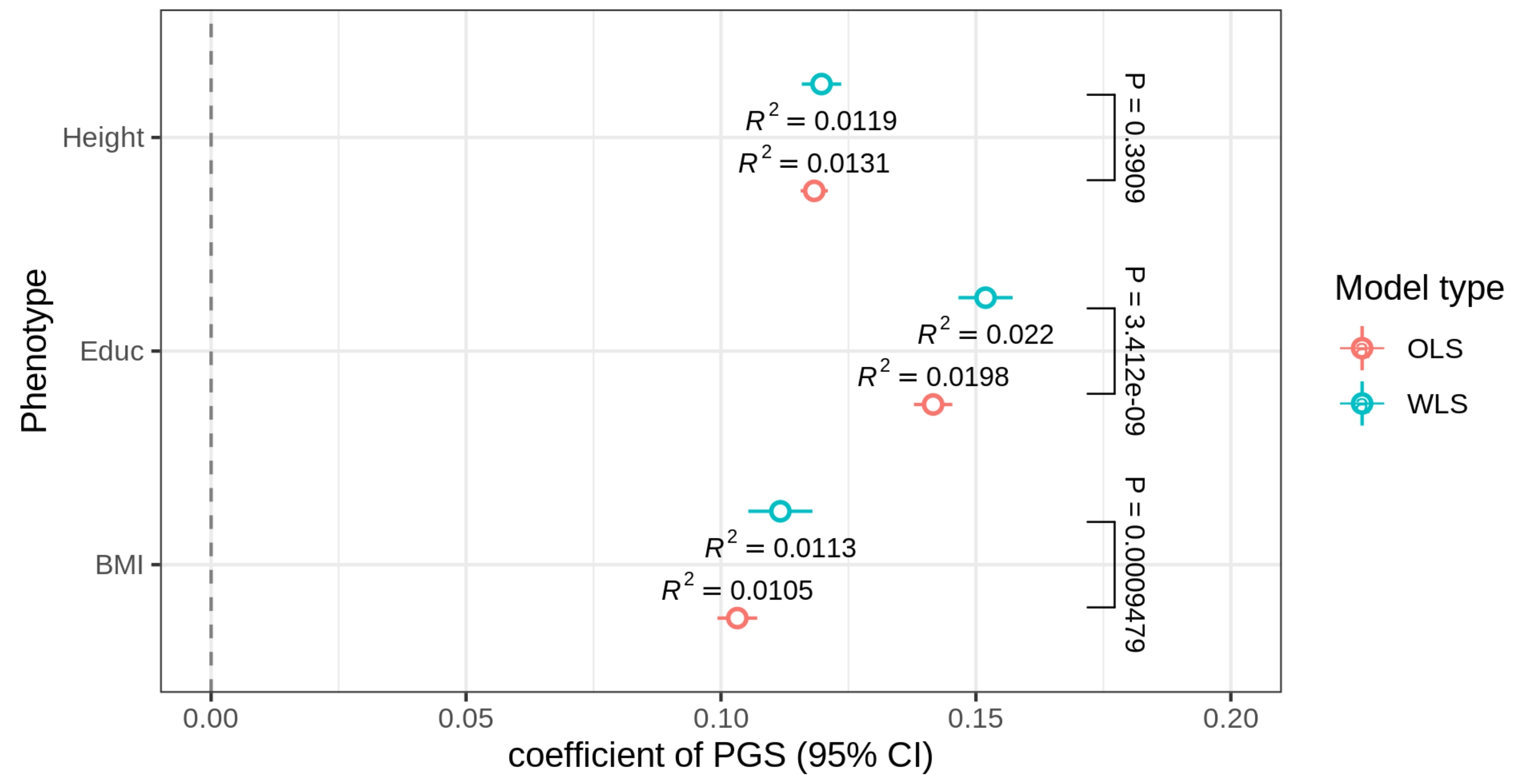


Figure 3

Results (3)

- Estimate SNP associations for 5,000 ``top hits'' for the three traits, as identified by recently published GWASs
- Unweighted and weighted SNP associations align closely (fig. 4)
- Weighted SNP associations that correct for selection show larger effect sizes on average (fig. 4)

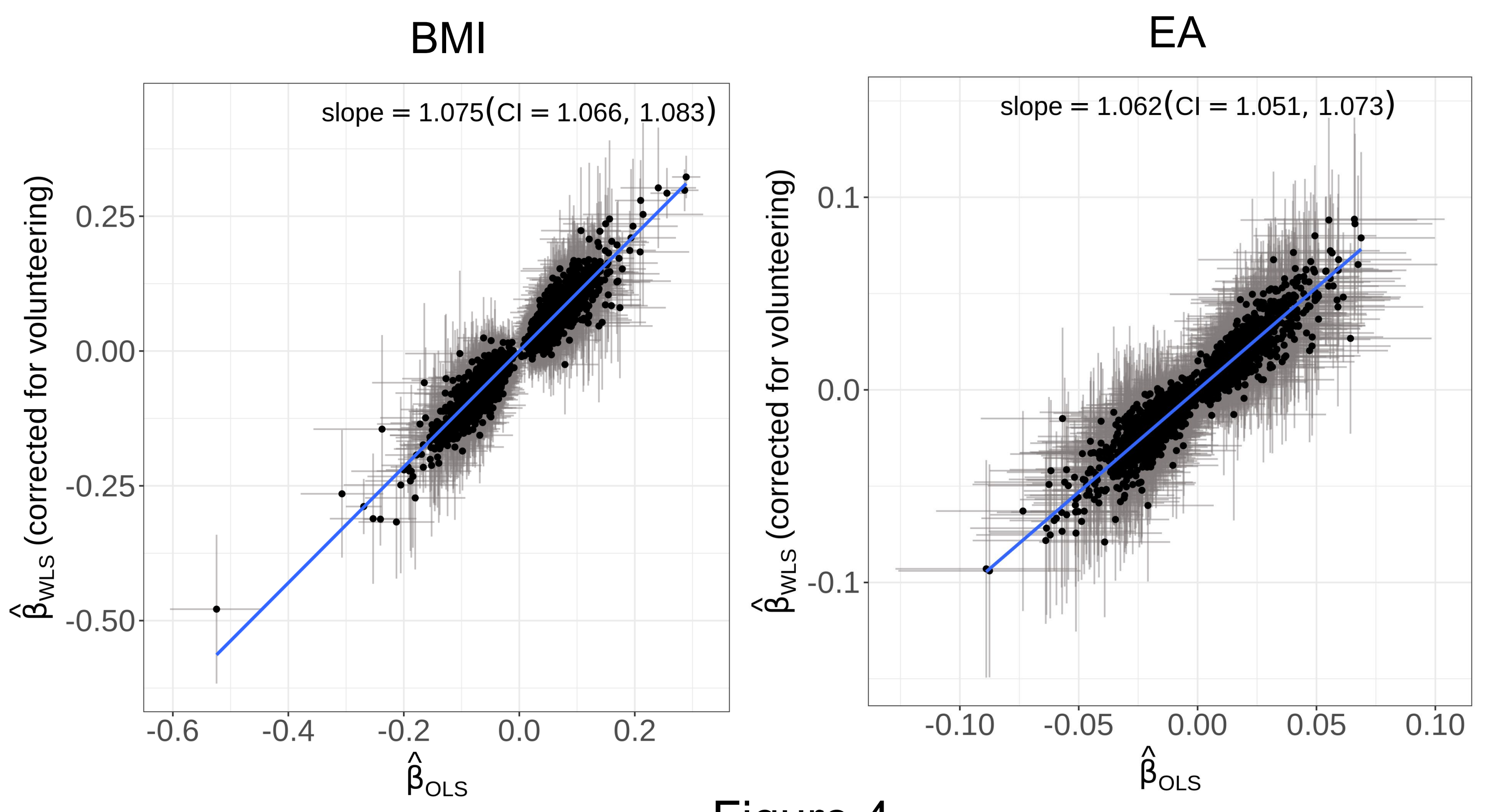


Figure 4

Conclusion

- Estimation of genetic associations in non-randomly selected samples results in volunteer bias
- Correcting for volunteer bias especially matters for traits with a large behavioral component (BMI, EA)