NAME OF PRESENTING AUTHOR: Sjoerd van Alten

EMAIL ADDRESS OF PRESENTING AUTHOR: s.j.d.van.alten@vu.nl

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

Sjoerd van Alten,[1,2] Benjamin W. Domingue,[3] Titus Galama,[1,2,4,5] and Andries Marees[1]

[1] School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[2] Tinbergen Institute, Amsterdam, Netherlands

[3] Stanford University, Stanford, California, USA

[4] Center for Economic and Social Research and Department of Economics, University of Southern California, Dornsife, Los Angeles, USA

[5] Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, Netherlands

KEYWORDS: Genome wide association studies, selection bias, UK Biobank

ABSTRACT:

Genome-wide association studies (GWASs) are almost always based on a non-random sample of the underlying population, as obtaining very large sample sizes, rather than ensuring such samples are representative, has been key to their success. Selection bias in estimated genetic associations, including how it varies across traits, is poorly understood. A sample of particular interest is the widely used UK Biobank (UKB). Individuals in the UKB are more likely to be female, higher educated, and older, due to its reliance on volunteering. Because of the need for very large samples, the UKB, by far the largest cohort, is included in almost all large GWASs. Further, UKB's subsample of genotyped siblings (UKBSIB) has become a crucial resource for estimating genetic effects free of environmental confounding. Using nationally representative UK Census microdata as a reference, we document substantial non-random selection into the UKB, and even stronger for UKBSIB. This non-random selection leads to significant selection bias in associations between various demographic and health-related traits when estimated in the UKB. We estimate probabilities of UKB participation for each UKB participant to estimate selection-corrected GWASs for multiple traits using (1) inverse probability weighting and (2) a Heckman two-step correction. We will assess whether selection-corrected GWAS results significantly differ from those of traditional GWASs and investigate which phenotypes are most affected, and why. Our results are useful for understanding whether a particular phenotype is prone to selection bias in GWAS and our correction method provides an alternative when population-representative cohorts are not available.

GRANT SUPPORT: We gratefully acknowledge financial support from NORFACE DIAL (462-16-100), the National Institute on Aging of the National Institutes of Health (RF1055654 and R56AG058726), and the Dutch National Science Foundation (016.VIDI.185.044).