A robust example of collider bias in a genetic association study

Felix R Day[1], Robert A Scott[1], Ken K. Ong[1] and John R.B. Perry[1#]

1. MRC Epidemiology Unit, Institute of Metabolic Science, Box 285, University of Cambridge, Hills Road, Cambridge, CB2 0QQ, UK

# correspondence: john.perry@mrc-epid.cam.ac.uk

A recent paper by Aschard *et al* described the potential for "collider bias" when adjusting for heritable covariates in genetic association studies[1]. However, in their examples the authors acknowledged that they could not exclude the possibility of a true biological explanation for the genetic association seen only in the adjusted model. Furthermore, the extent to which this bias could create a completely spurious genetic association, rather than just modify the magnitude of the effect,[2] remains unclear.

Collider bias describes the artificial association created between two uncorrelated exposures (*A* and *B*) when a shared outcome (*X*) is included in the model as a covariate (**Figure 1**). We sought to definitively illustrate collider bias by deliberately inducing it to generate a biologically implausible SNP-phenotype association. Both sex (A) and autosomal genetic determinants for adult height (B) have causal effects on height (*X*), but are themselves implausibly correlated. We theorized that collider bias would induce false-positive associations between (only) autosomal height-associated genetic variants and sex when adjusting for height as a covariate.

In a sample of 142,630 individuals of white European ancestry from the UK Biobank study[3], we performed a genome-wide association study for sex using a linear mixed model and applying standard quality control metrics. As expected, in univariate models, no SNP reached genome-wide significance and test statistics for the 697 previously identified height SNPs[4] conformed to a null distribution (P_{min}=1.4x10^{-3}, 27 SNPs P<0.05, ~35 by expected by chance). In contrast, when we repeated the analysis including height as a covariate, 222/697 height SNPs reached genome-wide significance for association with sex. Each height increasing allele exhibited the expected negative correlation with sex (i.e lower likelihood of being male), given the two causal exposures were aligned to be positively associated with height (**Figure 1**). This was exemplified by the three strongest signals in the reported GIANT height GWAS meta-analysis: *ZBTB38*-rs724016 (sex P_{unadj}=0.05, P_{adj}=7x10^{-90}), *GDF5*-rs143384 (P_{unadj}=0.13, P_{adj}=7x10^{-71}) and *HMGA2*-rs8756 (P_{unadj}=0.99, P_{adj}=3x10^{-34}), which all showed an apparent robust association with sex only in the height-adjusted model. Furthermore, amongst the 697 height-associated SNPs, their beta estimates for sex adjusted for height were almost perfectly correlated with their beta estimates for height adjusted for sex in this UK Biobank study sample (**Figure 1**).

In summary, we provide convincing evidence that adjusting for covariates can create apparently highly robust, but biologically spurious associations. The extent of this collider bias is almost perfectly inverse to the strength of the exposure-collider association. Consideration of causal inference modeling and unadjusted test statistics is therefore of great importance in the design and interpretation of genetic (and non-genetic) association studies.

Note: This article has been submitted as a "letter to the editor". Further information regarding statistical methods is available upon request.

Acknowledgements: This work was conducted using the UK Biobank resource.

References

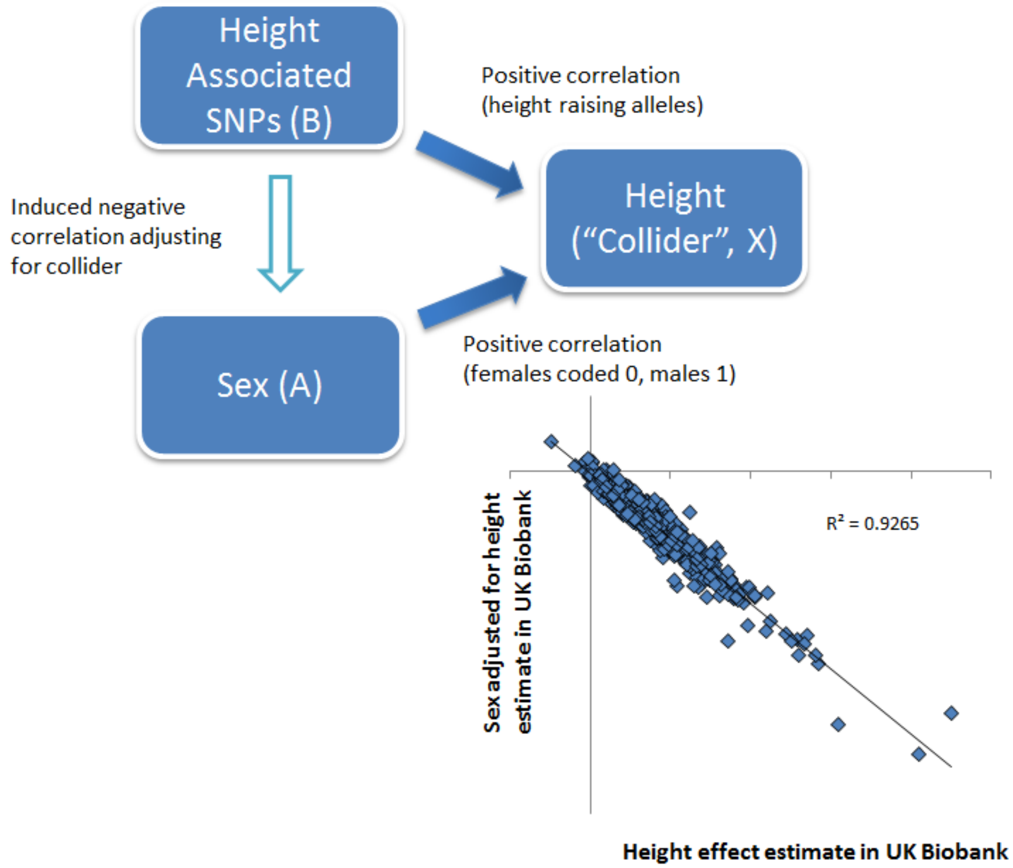
1. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96,** 329–39 (2015).

2. Yaghootkar, H. *et al.* Association analysis of 29,956 individuals confirms that a low-frequency variant at CCND2 halves the risk of type 2 diabetes by enhancing insulin secretion. *Diabetes* **64,** 2279–85 (2015).

3. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12,** e1001779 (2015).

4. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* (2014). doi:10.1038/ng.3097

Figure 1 | 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 697 previously identified genome-wide significant autosomal SNPs for height.


Height Associated SNPs (B)
Positive correlation (height raising alleles)
Height ("Collider", X)
Induced negative correlation adjusting for collider
Sex (A)
Positive correlation (females coded 0, males 1)
Sex adjusted for height estimate in UK Biobank
R^2 = 0.9265
Height effect estimate in UK Biobank