bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Collider Scope:

How selection bias can induce spurious associations

Marcus R. Munafò [1,2], Kate Tilling [1,3], Amy E. Taylor [1,2], David M. Evans [1,4], George Davey Smith [1,3]

1. MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom.
2. UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, United Kingdom.
3. School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom.
4. University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4102, Australia.

Corresponding author: Marcus R. Munafò, School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, United Kingdom. T: +44.117.9546841; F: +44.117.9288588; E: marcus.munafo@bristol.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Large-scale cross-sectional and cohort studies have transformed our understanding of the genetic and environmental determinants of health outcomes. However, the representativeness of these samples may be limited – either through selection into studies, or by attrition from studies over time. Here we explore the potential impact of this selection bias on results obtained from these studies. While it is acknowledged that selection bias will have a strong effect on representativeness and prevalence estimates, it is often assumed that it should not have a strong impact on estimates of associations. We argue that because selection can induce collider bias (which occurs when two variables independently influence a third variable, and that variable is conditioned upon), selection can lead to biased estimates of associations. In particular, selection related to phenotypes can bias associations with genetic variants associated with those phenotypes. In simulations, we show that even modest influences on selection into or attrition from a study can generate biased and potentially misleading estimates of both phenotypic and genotypic associations. Our results highlight the value of representative birth cohorts. Having DNA available on most participants at birth at least offers the possibility of investigating the extent to which polygenic scores predict subsequent participation, which in turn would enable sensitivity analyses of the extent to which bias might distort estimates.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Collider Scope:

How selection bias can induce spurious associations

Introduction

Understanding the impact of genetic and environmental factors on physical and mental health outcomes is critical if we are to develop effective preventive and treatment interventions. Large-scale cross-sectional and cohort studies provide an invaluable resource to support these efforts, in particular with respect to genetic influences, where the small effects associated with common genetic variants require very large samples to achieve adequate statistical power. However, achieving these very large sample sizes in population-based studies may come at the cost of representativeness – participants who volunteer to participate in studies may not be representative of the general population (1).

While some studies may be relatively representative at inception, through rigorous efforts to ensure representative recruitment (e.g., birth cohort studies), as they mature the likelihood is that attrition from the study will be non-random, so that the cohort becomes less representative of the general population as time goes on. There is already clear evidence from existing large-scale population studies that they are subject to a degree of selection bias. For example, higher genetic risk scores for schizophrenia are consistently associated with non-completion of questionnaires by study mothers and children, as well as non-attendance at data collection clinics, in the Avon Longitudinal Study of Parents and Children (ALSPAC) (2) (see Box 1).

Attrition from cohort studies may result in biased estimates of socioeconomic inequalities, and the degree of bias may worsen as participation rates decrease (3). However, it is often argued that representativeness is not necessary in studies of this kind (4-8), although this is not universally accepted (9). In particular, for genetic variants, where conventional confounding is low (10), it has been argued, even by those concerned about selection bias, that any problems associated with a lack of

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

representativeness may be modest (9, 11). Here we ask: What is the impact of selection bias on the results obtained from these studies?

Insert Box 1 about here.

Collider Bias

It is widely acknowledged selection bias will distort prevalence estimates. This can be clearly seen in differences between participants in the original ALSPAC sample and those that attended later clinics (see Box 1), as well as in the UK Biobank study relative to the general population (see Box 2). However, it is often assumed that whilst selection bias will have a strong effect on representativeness and prevalence estimates, it should not have a strong impact on observed associations (4). This overlooks the fact that selection bias can in turn induce collider bias (see Figure 1), which can lead to spurious observational and genetic associations.

Insert Figure 1 and Box 2 about here.

Collider bias occurs when two variables (*X* and *Y*) independently cause a third variable (*Z*). In this situation, *Z* is a collider, and statistical adjustment for *Z* will bias the estimated causal association of *X* (exposure) on *Y* (outcome) (see Figure 2). Statistical adjustment of the XY association for a variable *Z* is equivalent to observing this association in a sub-population where all individuals share the same value of *Z* (1, 12). Hence if both X and Y cause participation in a study (*Z*), then investigating associations in the selected sample (i.e., with Z = 1, indicating participation in) is equivalent to conditioning on *Z*, which in turn may induce collider bias.

Insert Figure 2 about here.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Put simply, statistical control is not equivalent to experimental control (1), and so sample selection can induce spurious associations between variables that influence participation or retention in a study, when no such association exists in the wider general population from which the sample is drawn. Alternatively, if two variables are correlated in the wider population, and both cause selection, then estimated correlation in the selected sample may be biased. Moreover, this selection bias will apply to the genetic correlates (or other ancestors) of these variables, unless the phenotypes are also controlled for. So if genes Gx and Gy cause X (exposure) and Y (outcome) respectively, then in the selected sample Gx will appear to be associated with Y (unless X is also controlled for). More complex situations can also give rise to collider bias, such as when the outcome (Y) doesn't directly *cause* selection into the study (i.e., it is a downstream consequence of something else that *is* causing selection into the study). However, it is necessary that the exposure (X) either directly or indirectly (such as in the situation described above) causes selection into the study.

In other words, traits that are entirely unrelated in the general population may appear to be correlated in selected samples, if both traits influence participation (and therefore contribute to selection), as a result of implicitly conditioning on their common effect (1, 13). There are exceptions to this depending on the distribution of the outcome and the parametric analysis model used. For example, if the outcome (Y) is a binary phenotype, and logistic regression is used, then the odds ratio for the association between the SNP and outcome may be unbiased even when the outcome causes selection (14).

We have previously argued that these effects may be greater in case-control studies than prospective studies, and that since genetic associations have been similar across study designs, the impact of selection bias may in fact be modest (11). We have also previously argued that because conventional confounding is typically low for single genetic variants, problems of selection bias will be less in this context

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(9). However, given the rapid growth in studies using data from highly selected samples such as UK Biobank, and the use of genetic scores rather than single genetic variants, we revisited this question, and used simulation to explore the potential impact of even relatively weak effects on participation. Given empirical evidence of selection in cross-sectional and cohort studies, what is the potential impact of this on observed phenotypic and genotypic associations?

Simulations

We simulated data on an allele score, a phenotype and an outcome, where both the phenotype and outcome influence selection into the study, but there was no association between the allele score and the outcome in the underlying population (see Figure 2). The simulation scenario was based on the UK Biobank. All variables were Normally distributed, with standard deviation of 1, and the sample size of the underlying complete population was 9,000,000. We assumed that phenotype and outcome had independent effects (i.e., no interaction on the additive scale) on the odds of selection into the sample, and for convenience we set these effects to be equal, and examined a weak association (OR of 1.2 for missingness for a 1 SD increase in phenotype/outcome) and two stronger associations (ORs of 1.5 and 1.8). These odds ratios are similar to estimates of the likelihood of participation in UK Biobank for individuals with any educational or vocational qualifications and for non–smokers, respectively (see Box 2), and indicate a difference in mean phenotype/outcome of 0.2 SD, 0.4 SD and 0.6 SD between those participating and those not participating. We varied the correlation between the allele score and the phenotype (between r = 0.05 and r = 0.30) to simulate genetic instruments explaining between 0.25% and 9% of the variance in phenotypes. These values are in the typical range for the association between common genetic variants, or polygenic risk scores comprising multiple common variants, and complex phenotypes. For example, the rs16969968 variant accounts for approximately 1% of the phenotypic variance in

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cigarette consumption (15), while the polygenic risk score for height captures approximately 9% of phenotypic variance (16). We controlled the baseline risk of selection into the sample, resulting in a selected sample of approximately 500,000 people. The analysis was an unadjusted regression of outcome on allele score not adjusting for the phenotype). In the whole population, the regression coefficient for outcome on allele score is zero, and the confidence interval contains zero 95% of the time. We simulated each scenario 100 times.

The results of this simulation are shown in Table 1, and indicate that the effects of selection bias are strongest for stronger independent selection effects, and also where the allele score is more strongly associated with the phenotype. However, even for moderate associations between missingness and both phenotype and outcome (OR = 1.5 for both phenotype and outcome) and between allele score and phenotype (r = 0.1, 1% variance explained by allele score) the confidence intervals contains zero only 89% of the time, and this continues to decrease with both greater strength of association between phenotype, outcome and missingness, and stronger association between allele score and phenotype.

Insert Table 1 about here.

Conclusions

Our results indicate the potential for unrepresentative samples to generate biased and potentially misleading estimates of both phenotypic and genotypic associations. In particular, when polygenic scores associated with a phenotype that combine many genetic variants are used, association between the phenotype and participation will cause the score to be more strongly related to participation than each individual variant is. This, in turn, can potentially lead to serious bias. For this reason, studies using polygenic scores, genome-wide allelic scores (17), and/or

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

whole-genome genetic correlations (18, 19) in highly unrepresentative studies are most at risk of producing biased and potentially misleading results.

The magnitude of effects we observed in our simulations, based on credible estimates of associations between both a phenotype or outcome and missingness, and between a polygenic score and a phenotype, are comparable with many reported associations derived from large but unrepresentative samples, such as between personality and cognitive function, and a range of physical and mental health outcomes (20, 21), and between chronotype (i.e., “morningness”) and years of education (22). An appreciation of the potential impact of selection bias may also resolve inconsistencies in the literature, and help to explain apparently paradoxical findings. For example, genetic correlations between cognitive ability and a range of psychiatric disorders have been reported to differ in childhood and older age (23). One possible interpretation is that this is due to age-dependent pleiotropy, but another is that this is an artefact of different selection bias pressures at different ages. An example serves to illustrate this. Polygenic risk scores that maximally capture schizophrenia liability are associated with increased psychotic experiences in ALSPAC participants, but scores that use more stringent thresholds for including genetic variants are associated with *reduced* psychotic experiences (24). Since missing data are likely to be greater for participants who report psychotic experiences, as well as for those at higher genetic risk of a psychotic disorder, , psychotic experiences may be relatively under-represented in participants with higher genetic risk, compared to those with lower genetic risk (24).

A related issue is the use of case-control studies to examine associations with “secondary” outcomes – that is, phenotypes other than the case/control outcome (25, 26). In such studies, the association between genotype and secondary phenotype will be biased if both genotype and secondary phenotype are associated with case-control status. Case-control studies condition on case-control status, and thus again collider bias can bias the association between genotype and secondary

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

phenotype. Various methods have been proposed to overcome this bias, including maximum likelihood and inverse probability weighting. This latter method requires some knowledge about the prevalence of case/control status in the underlying population, or the assumption that the disease is rare (25, 26).

We have discussed one important way in which selection into or out of a study can induce collider bias and spurious associations. There are other ways in which ascertainment can generate biases (27). For example, Figure 3 (panel B) shows a situation in which entry into a study is conditional upon the value of the phenotype (but not the outcome of interest) and where the phenotype does not cause the outcome, but the phenotype and outcome are correlated in unselected samples (i.e., due to genetic and/or environmental factors U). In this situation, collider bias occurs because conditioning on selection induces an association between SNPs related to the phenotype and the polygenic and/or environmental factors that influence the outcome. Therefore SNPs that cause the phenotype only (i.e. do not in truth cause the outcome), may now show spurious relationships with the outcome variable. Figure 3 (panels C to E) also shows examples where selection will bias the estimation of the causal effects of SNPs on the outcome. In these examples, SNPs that do cause the outcome directly via the phenotype will either show increased or decreased association in the selected sample, depending on the underlying genetic and environmental aetiology of both traits. Other, more complex, situations can also lead to selection bias – we have not attempted to outline every possible case here. Algorithms for deciding whether a given causal analysis is biased by selection have been described (28), and could be used to decide whether bias is likely in a given case.

Insert Figure 3 about here.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our results highlight the value of representative birth cohorts. Having DNA available on all participants at birth at least offers the possibility of investigating the extent to which polygenic scores predict subsequent participation. Without this knowledge, studies in large, unrepresentative samples run the risk of providing biased and misleading results. In our opinion these important caveats should be borne in mind when interpreting the results of such studies.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

MRM is a member of the UK Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. This work was supported in part by the Medical Research Council and the University of Bristol (MC_UU_12013/1, MC_UU_12013/4, MC_UU_12013/6, MC_UU_12013/9).

Conflict of Interest

There are no conflicts of interest to declare.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1 Lee, J.J. (2012) Correlation and causation in the study of personality. *European Journal of Personality*, **26**, 372-390.

2 Martin, J., Tilling, K., Hubbard, L., Stergiakouli, E., Thapar, A., Davey Smith, G., O'Donovan, M.C. and Zammit, S. (2016) Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *American Journal of Epidemiology*.

3 Howe, L.D., Tilling, K., Galobardes, B. and Lawlor, D.A. (2013) Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. *Epidemiology*, **24**, 1-9.

4 Rothman, K.J., Gallacher, J.E. and Hatch, E.E. (2013) Why representativeness should be avoided. *International journal of epidemiology*, **42**, 1012-1014.

5 Elwood, J.M. (2013) Commentary: On representativeness. *International Journal of Epidemiology*, **42**, 1014-1015.

6 Nohr, E.A. and Olsen, J. (2013) Commentary: Epidemiologists have debated representativeness for more than 40 years--has the time come to move on? *International Journal of Epidemiology*, **42**, 1016-1017.

7 Richiardi, L., Pizzi, C. and Pearce, N. (2013) Commentary: Representativeness is usually not necessary and often should be avoided. *International Journal of Epidemiology*, **42**, 1018-1022.

8 Rothman, K.J., Gallacher, J.E. and Hatch, E.E. (2013) Rebuttal: When it comes to scientific inference, sometimes a cigar is just a cigar. *International Journal of Epidemiology*, **42**, 1026-1028.

9 Ebrahim, S. and Davey Smith, G. (2013) Commentary: Should we always deliberately be non-representative? *International Journal of Epidemiology*, **42**, 1022-1026.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10 Davey Smith, G., Lawlor, D.A., Harbord, R., Timpson, N., Day, I. and Ebrahim, S. (2007) Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine*, **4**, e352.

11 Davey Smith, G. (2012) The Wright stuff: Genes in the interrogation of correlation and causation. *European Journal of Personality*, **26**, 395-397.

12 Hernan, M.A., Hernandez-Diaz, S. and Robins, J.M. (2004) A structural approach to selection bias. *Epidemiology*, **15**, 615-625.

13 Asendorpf, J.B. (2012) Bias due to controlling a collider: A potentially important issue for personality research. *European Journal of Personality*, **26**, 391-413.

14 Bartlett, J.W., Harel, O. and Carpenter, J.R. (2015) Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology*, **182**, 730-736.

15 Ware, J.J., van den Bree, M.B. and Munafo, M.R. (2011) Association of the CHRNA5-A3-B4 gene cluster with heaviness of smoking: a meta-analysis. *Nicotine & Tobacco Research*, **13**, 1167-1175.

16 Zhang, B., Shu, X.O., Delahanty, R.J., Zeng, C., Michailidou, K., Bolla, M.K., Wang, Q., Dennis, J., Wen, W., Long, J. *et al.* (2015) Height and breast cancer risk: Evidence from prospective studies and Mendelian randomization. *Journal of the National Cancer Institute*, **107**.

17 Evans, D.M., Visscher, P.M. and Wray, N.R. (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, **18**, 3525-3531.

18 Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. and Wray, N.R. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540-2542.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19 Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen, C., Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, Duncan, L. *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, **47**, 1236-1241.

20 Gale, C.R., Hagenaars, S.P., Davies, G., Hill, W.D., Liewald, D.C., Cullen, B., Penninx, B.W., International Consortium for Blood Pressure GWAS, CHARGE Consortium Ageing and Longevity Group, Boomsma, D.I. *et al.* (2016) Pleiotropy between neuroticism and physical and mental health: findings from 108 038 men and women in UK Biobank. *Translational Psychiatry*, **6**, e791.

21 Hagenaars, S.P., Harris, S.E., Davies, G., Hill, W.D., Liewald, D.C., Ritchie, S.J., Marioni, R.E., Fawns-Ritchie, C., Cullen, B., Malik, R. *et al.* (2016) Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Molecular Psychiatry*.

22 Lane, J.M., Vlasac, I., Anderson, S.G., Kyle, S.D., Dixon, W.G., Bechtold, D.A., Gill, S., Little, M.A., Luik, A., Loudon, A. *et al.* (2016) Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nature Communications*, **7**, 10889.

23 Hill, W.D., Davies, G., Charge Cognitive Working Group, Liewald, D.C., McIntosh, A.M. and Deary, I.J. (2015) Age-dependent pleiotropy between general cognitive function and major psychiatric disorders. *Biological Psychiatry*.

24 Jones, H.J., Stergiakouli, E., Tansey, K.E., Hubbard, L., Heron, J., Cannon, M., Holmans, P., Lewis, G., Linden, D.E., Jones, P.B. *et al.* (2016) Phenotypic manifestation of genetic risk for schizophrenia during adolescence in the general population. *JAMA Psychiatry*, **73**, 221-228.

25 Xing, C., J, M.M., Dupuis, J., Adrienne Cupples, L., J, B.M., Lin, X. and A, S.A. (2016) Robust analysis of secondary phenotypes in case-control genetic association studies. *Statistics in Medicine*.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26 Song, X., Ionita-Laza, I., Liu, M., Reibman, J. and We, Y. (2016) A general and robust framework for secondary traits analysis. *Genetics*, **202**, 1329-1343.

27 Elwert, F. and Winship, C. (2014) Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, **40**, 31-53.

28 Daniel, R.M., Kenward, M.G., Cousens, S.N. and De Stavola, B.L. (2012) Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, **21**, 243-256.

29 Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S. and Davey Smith, G. (2013) Cohort Profile: The 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*, **42**, 111-127.

30 Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A. *et al.* (2013) Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology*, **42**, 97-110.

31 Shweikh, Y., Ko, F., Chan, M.P., Patel, P.J., Muthy, Z., Khaw, P.T., Yip, J., Strouthidis, N., Foster, P.J., Eye, U.K.B. *et al.* (2015) Measures of socioeconomic status and self-reported glaucoma in the U.K. Biobank cohort. *Eye*, **29**, 1360-1367.

32 Office for National Statistics. (2014). Adult Smoking Habits in Great Britain: 2014.

33 Office for National Statistics. (2011). Census: 2011.

34 Ganna, A. and Ingelsson, E. (2015) 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. *Lancet*, **386**, 533-540.

35 Wain, L.V., Shrine, N., Miller, S., Jackson, V.E., Ntalla, I., Soler Artigas, M., Billington, C.K., Kheirallah, A.K., Allen, R., Cook, J.P. *et al.* (2015) Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respiratory Medicine*, **3**, 769-781.

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Box 1. The Avon Longitudinal Study of Parents and Children.

Birth cohort studies are also not immune to problems of selection bias, where retention in the study may be related to a variety of participant characteristics. The Avon Longitudinal Study of Parents and Children (ALSPAC) recruited pregnant women living in the administrative county of Avon with expected delivery dates between 1st April 1991 and 31st December 1992. These women, their partners and their offspring have been followed up ever since via questionnaires and clinics. ALSPAC originally captured data on 14,541 pregnancies (75% of eligible women) (29, 30), but inevitably retention in subsequent data collection sweeps (postal questionnaires and clinic assessments) was less than 100%. We see that higher body mass index (BMI) is associated with lower odds of subsequent retention in both mothers (N = 11,319, OR per SD increase in BMI 0.85, 95% CI 0.81 to 0.88), for retention between 2008 and 2011 using pre-pregnancy BMI as a predictor, and offspring (N = 7,954, OR 0.91, 95% CI 0.87 to 0.96), for retention at age 18 using BMI at age 7 as a predictor. Similarly, among smoking mothers in ALSPAC, heaviness of smoking is associated with lower odds of retention (N = 3,534, OR per additional cigarette smoked per day just prior to pregnancy 0.97, 95% CI 0.96 to 0.98). If low BMI and maternal non-smoking are both related to continuing participation in ALSPAC, this would tend to lead to the association between BMI and maternal smoking being negatively biased (i.e., we would expect to see a more negative association between genetic variants positively associated with smoking and BMI in ALSPAC than in the true underlying population).

bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Box 2. UK Biobank.

The UK Biobank is a cross sectional study, which recruited over 500,000 individuals aged between 40 and 69 years between 2006 and 2010 (see http://www.ukbiobank.ac.uk/). Individuals in this age group living within a 25 mile radius of any of the 22 assessment centres across the UK were identified from NHS patient registers (31). In total, around 9 million individuals were invited to participate. However, UK Biobank was only able to achieve a 5% response rate (~500,000 participants recruited from ~9,000,000 invited, personal communication, UK Biobank, 8th July 2016), and the resulting sample is not representative of the UK population as a whole. For example, the proportion of current smokers is relatively low in UK Biobank (11% vs 19% in the general population, equivalent to an OR of 1.89) (32), as is the proportion with no qualifications (18% vs 25%, equivalent to an OR of 1.50) (33). Unsurprisingly, therefore, participants in UK Biobank have far lower rates of 5-year mortality than the UK population as a whole (34). Clearly, agreeing to take part in UK Biobank study is associated with a number of characteristics that will reflect, for example, health status and social position. If non-smoking and having qualifications are both causally related to participation in UK Biobank, we would expect the association between smoking and having qualifications to be positively biased (i.e., we would expect to see a more positive association between genetic variants positively associated with smoking and whether participants had educational qualifications in UK Biobank than in the true population). The problem is possibly compounded in genetic studies using the first release of genomewide association data in UK Biobank, which used two genotyping arrays, one of which was applied to a nested case-control study of smoking and lung function (UK BiLEVE) (35). The first release genetic data are therefore further subject to selection bias relative to UK Biobank as a whole (although this will no longer be the case when the full release of genomewide association data becomes available).

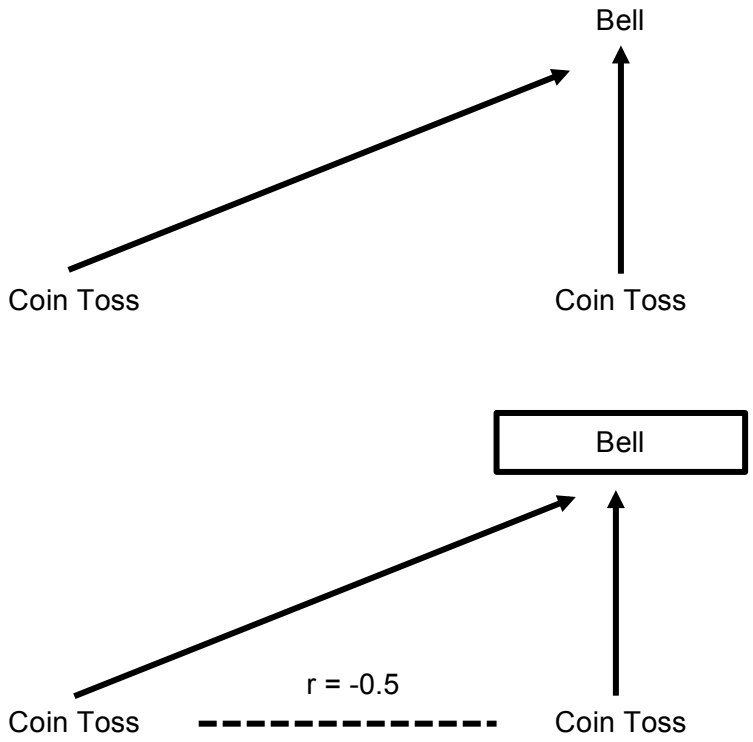
bioRxiv preprint doi: https://doi.org/10.1101/079707; this version posted October 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Results of simulation study showing the selection bias in estimating an association which is null in the underlying population.

Simulation settings		Results – association between allele score and outcome		
Association between missingness and both phenotype and outcome (OR)	Association between allele score and phenotype (r)	Mean regression coefficient (SD)	Mean z-score (SD)	Number of 95% CIs containing zero
OR = 1.8	0.05 (0.25% variance)	-0.001 (0.001)	-1.04 (1.00)	83
	0.10 (1.00% variance)	-0.003 (0.001)	-2.06 (0.98)	45
	0.15 (2.25% variance)	-0.004 (0.001)	-3.07 (0.98)	9
	0.20 (4.00% variance)	-0.006 (0.001)	-4.10 (0.98)	0
	0.30 (9.00% variance)	-0.008 (0.001)	-6.18 (1.06)	0
OR = 1.5	0.05 (0.25% variance)	-0.001 (0.001)	-0.42 (0.95)	94
	0.10 (1.00% variance)	-0.001 (0.001)	-0.80 (0.96)	89
	0.15 (2.25% variance)	-0.001 (0.001)	-1.22 (0.96)	77
	0.20 (4.00% variance)	-0.002 (0.001)	-1.64 (0.97)	61
	0.30 (9.00% variance)	-0.003 (0.001)	-2.44 (0.94)	35
OR=1.2	0.05 (0.25% variance)	-0.0002 (0.001)	-0.16 (0.92)	97
	0.10 (1.00% variance)	-0.0003 (0.001)	-0.25 (0.94)	97
	0.15 (2.25% variance)	-0.0005 (0.001)	-0.38 (0.95)	93
	0.20 (4.00% variance)	-0.0006 (0.001)	-0.47 (0.95)	91
	0.30 (9.00% variance)	-0.0009 (0.001)	-0.66 (0.96)	89

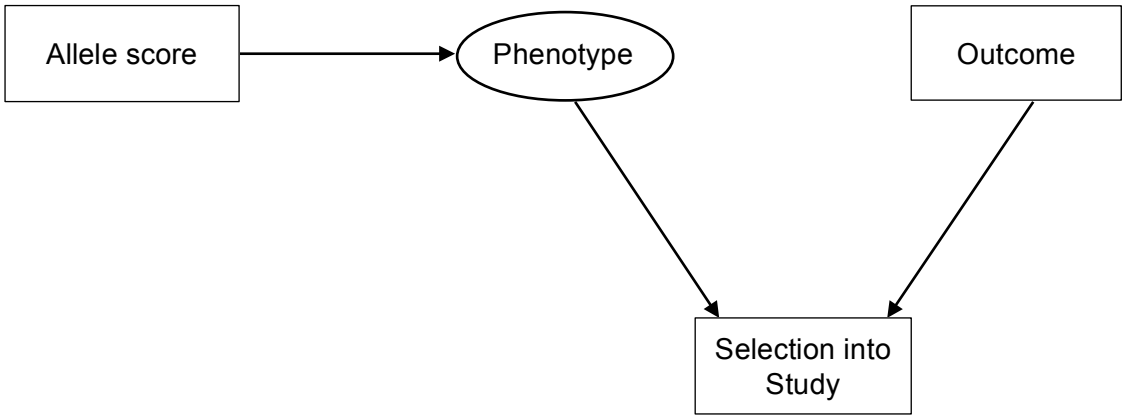
OR: odds ratio; r: correlation coefficient; SD: standard deviation; CI: confidence interval. Each scenario was simulated 100 times.

Figure 1. Illustration of collider bias.



The basic premise of collider bias is shown. In this example, a bell is sounded whenever either coin come up 'heads'. The result of one coin toss is independent of the other. However, if we hear the bell ring (i.e., we condition on the bell ringing), then if you see a tail on one coin you know there must be a head on the other – the two coin results are no longer independent and a spurious inverse correlation has been induced. Reproduced from Gage SH, Davey Smith G, Ware JJ, Flint J, Munafò MR (2016) G = E: What GWAS Can Tell Us about the Environment. PLoS Genet 12(2): e1005765. doi:10.1371/journal.pgen.1005765

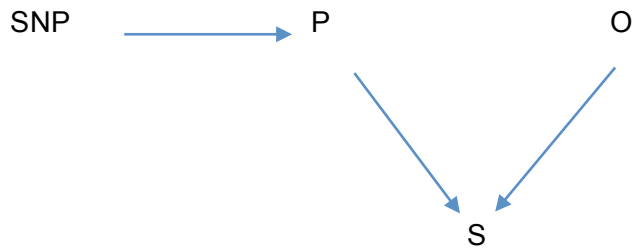
Figure 2. Illustration of selection bias simulation.



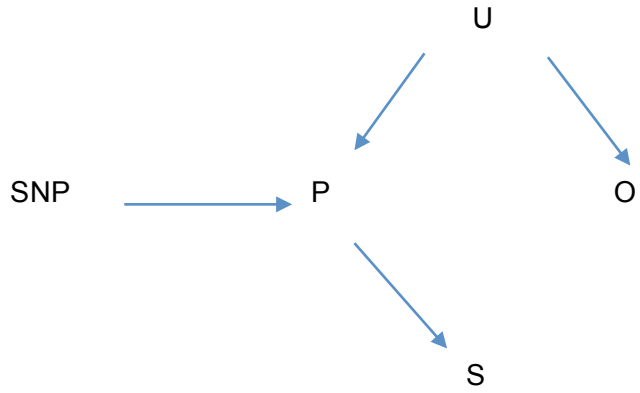
In the entire population there is no association between allele score and outcome. Selection into the study (either through voluntary participation at baseline, or attrition over time) induces an association between allele score and outcome (collider bias).

Figure 3. Scenarios where selection bias would occur.

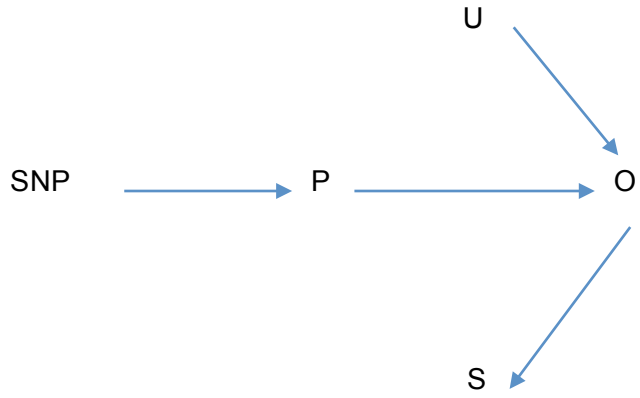
A. In truth the SNP is not causally associated with the outcome; selection will induce an association (which could be positive or negative).



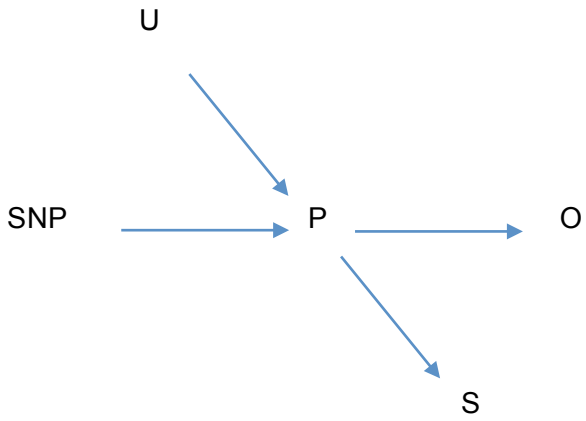
B. In truth the SNP is not causally associated with the outcome; selection will induce an association (which could be positive or negative).



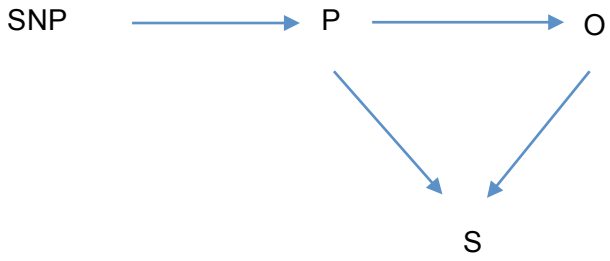
C. In truth the SNP is causally associated with the outcome; selection could make this larger or attenuate it.



D In truth the SNP is causally associated with the outcome; selection could make this larger or attenuate it.



E. In truth the SNP is causally associated with the outcome; selection will bias this association (which could be positive or negative).



SNP: single nucleotide polymorphism; P: Phenotype; O: Outcome; S: Selection.