Original Research Article

Title: Bias from competing risk before recruitment in Mendelian Randomization studies of conditions with shared etiology

Schooling CM[1,2], Lopez P[1], Yang Z,[2] Au Yeung SL[2], Huang JV[2]

[1]CUNY Graduate School of Public Health and Health Policy, New York, NY, USA

[2]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

Emails

C M Schooling: maryschooling@sph.cuny.edu; cms1@hku.hk

Corresponding author

C M Schooling, PhD

55 West 125th St, New York, NY 10027, USA

Phone 646 364-9519

Fax 212 396 7644

Conflicts of interest: There is no conflict of interest

Sources of financial support: None

Data & Code: The data used here is publicly available, the code is available on request.

Acknowledgements: None

Abstract: 200

Word count: 2818

Figures: 2

Tables: 1

Appendix: 1

Supplementary Tables: 1

Abstract

Mendelian randomization, i.e., instrumental variable analysis with genetic instruments, can foreshadow findings from randomized controlled trials quickly and cheaply even when no study measuring both exposure and outcome exists. Mendelian randomization studies occasionally generate paradoxical findings substantially different from trial estimates. Despite robustness to confounding, Mendelian randomization studies are, like all observational studies, open to selection bias. Any gap between randomization (at conception) and recruitment may mean the study recruits survivors not only of the genetically predicted exposure but also of other diseases that share etiology with the outcome, i.e., competing risks. Many chronic conditions share etiology. Here we show theoretically, empirically and analytically, that in survivors of a genetically predicted exposure, failure to account for competing risk of the outcome due to shared etiology can generate paradoxical Mendelian randomization estimates, particularly because genetic associations tend to be small and so easily reversed. The resulting bias can be conceptualized as dependent on the proportion of the genetically exposed who have already died of the exposure and of potential cases of the outcome who have already died of competing risk. Interpreting Mendelian randomization estimates within the natural history of diseases may help but methods that obviate this bias are required urgently.

Keywords: selection bias, competing risk, Mendelian randomization, instrumental variable analysis, shared etiology

Mendelian Randomization (MR), i.e., instrumental variable analysis with genetic instruments, is an increasingly popular and influential analytic technique [1, 2], which can be used to investigate causal effects even when no study including both exposure and outcome of interest exists. Invaluably, MR studies have provided estimates more consistent with results from randomized controlled trials (RCTs) than conventional observational studies, even foreshadowing the results of major trials [3]. MR studies are often presented as observational studies analogous to RCTs [4, 5] because they take advantage of the random assortment of genetic material at conception, while observational studies are open to biases from confounding and selection bias [6]. Instrumental variable analysis is described in health research as addressing confounding [7, 8], i.e., bias from common causes of exposure and outcome [6]. MR is currently described as "less likely to be affected by confounding or reverse causation than conventional observational studies" [2].

MR was originally thought to be less open to selection bias than conventional observation studies [9]. Selection bias is now increasingly widely recognized as a limitation of MR [10-20] which may violate the instrumental variable assumptions. Sources of potential selection bias in MR have been specifically identified as selecting an unrepresentative sample [10, 12, 13], attrition from an initially representative sample, such as a birth cohort [13], and selecting a sample strongly on surviving the exposure [11] or genotype of interest [16, 21]. What has not explicitly been considered is selecting the sample on surviving the genotype/exposure of interest in the presence of competing risk of the outcome. MR studies are particularly vulnerable to sample selection on survival because of the time lag between randomization (at conception) and typical recruitment into studies of major diseases in middle- to old-age. MR studies also often concern major causes of death thought to share considerable etiology. For example, lipids, blood pressure, diabetes, lifestyle (smoking, diet, physical activity, sleep) and socio-economic position cause both ischemic heart disease (IHD) and ischemic stroke, with death from IHD typically occurring at younger ages than death from stroke [22]. As a result, a study of the association of lipids with stroke among the living will automatically select on surviving unhealthy lipids and on surviving competing risk of

prior death from IHD due to their shared etiology. Here, we consider theoretically how such situations relate to the instrumental variable assumptions, provide empirical illustrative examples, explain the relevance to paradoxical MR findings and to current areas of research, explain the implications for MR and provide a simple analytic means of assessing such potential bias.

Potential biasing pathways due to selective survival

Figure 1a shows the directed acyclic graph for MR illustrating the instrumental variable assumptions typically referred to as relevance, independence and exclusion-restriction. Relevance is explicitly indicated by the arrow from instrument to exposure. Independence is implicitly indicated by the lack of an arrow from confounders of exposure on outcome to instrument. Exclusion-restriction is implicitly indicated by the lack of arrows linking instrument to outcome, sometimes illustrated as no arrow from instrument to outcome indicating no pleiotropy [23-26] (Figure 1b). Figure 1c shows selection on survival of both instrument and common causes of the outcome (U_2) [12, 19], which also violates the exclusion restriction assumption, particularly when stated as *"every unblocked path connecting instrument and outcome must contain an arrow pointing into the exposure"* [27]. Figure 1d explicitly shows survival on instrument, and another disease (Y_2) sharing etiology (U_2) with the outcome (Y_1). Figure 1e shows the exclusion restriction assumption with both no pleiotropy and no selection bias from competing risk made explicit.

Notably, Figures 1c and 1d are very similar in structure to a well-known example of selection bias which occurs when conditioning on an intermediate reverses the direction of effect: the "birth weight" paradox [28]. In the birth weight paradox, the positive association of maternal smoking with infant death becomes inverse after adjusting for birth weight, likely because birth weight is affected by maternal smoking, and by a common cause of birth weight and infant death, i.e., infant defects, which was not included in the analysis [28]. In the birth weight paradox adjusting for all common causes of birth weight and survival, i.e. birth defects, should remove the bias [28] because it blocks the path from survival to outcome.

Common study designs, such as cross-sectional, case-control or cohort, usually recruit from those currently alive. In such studies selection bias from competing risk before recruitment would be particularly marked when assessing the effect of a harmful exposure on an outcome that shares etiology with other diseases that commonly cause death at earlier ages than the outcome of interest (Figure 1d). Paradoxical reversals of effects in observational studies of older people or sick people are well known as examples of selection bias, such as smoking being inversely associated with dementia [29], which has also occurred in an MR studies [30]. MR studies also consistently suggest no effect of body mass index (BMI) on stroke [31], which could be an overlooked etiological difference between IHD and stroke or could be due to bias from inadvertently selecting the sample on surviving both genetically predicted BMI and IHD because of the many shared causes of IHD and stroke (Figure 2a), when IHD typically causes death before (at younger ages) than stroke [22]. Specifically, genetically high BMI may appear unrelated to stroke because before recruitment into a stroke case-control study some people die from genetically high BMI and some people die of IHD, which shares causes with stroke. People can only die once so that leaves a shortage of people available to recruit with genetically high BMI and stroke, so genetically high BMI appears unrelated to stroke. To put it another way people remaining alive and so available to recruit have survived high BMI or other IHD risk factors. If a survivor has high BMI they are less likely to have other IHD risk factors, and conversely if a survivor has other IHD risk factors they are less likely to have high BMI, meaning the true effect of high BMI on stroke cannot be observed and high BMI may appear to have a neutral or even protective effect. Below we provide an illustrative example.

Illustrative example

Statins and PCSK9 inhibitors are well-established interventions for cardiovascular disease, which reduce low density lipoprotein (LDL)-cholesterol, IHD [32-34], stroke [32-34] and atrial fibrillation (AF) [35]. IHD, stroke and AF also share major causes independent of LDL-cholesterol, such as blood pressure [36, 37], lifestyle and socio-economic position, but IHD may typically occur first [22], and AF may be a

consequence of IHD. Statins also appear to have a greater effect on overall survival than PCSK9 inhibitors [33, 35, 38, 39]. As such, given Figure 2b, greater bias would be expected when using MR to assess effects of harmful exposures on stroke and AF than on IHD, for studies with participants recruited at older ages and possibly more for genetically predicted statins than PCSK9 inhibitors.

We obtained genetically predicted LDL based on well-established genetic variants predicting the effects of statins (rs12916, rs17238484, rs5909, rs2303152, rs10066707 and, rs2006760) and PCSK9 inhibitors (rs11206510, rs2479409, rs2149041, rs2479394, rs10888897, rs7552841 and rs562556) [40]. We applied these variants to major genome wide association studies (GWAS) in people largely of European descent of IHD (CARDIoGRAMplusC4D 1000 Genomes) [41], ischemic stroke (MEGASTROKE) [42] and AF (Nielsen et al) [43] and to the UK Biobank summary statistics for IHD and ischemic stroke [44], but not AF because the GWAS includes relevant data from the UK Biobank [43]. Appendix Table 1 gives descriptive information about these GWAS. We obtained inverse variance weighted MR estimates with multiplicative random effects taking into account any correlations between genetic variants for European populations obtained from LD-Link (https://ldlink.nci.nih.gov) using the Mendelianrandomization R package.

Table 1 shows the MR associations of genetically predicted lower LDL-cholesterol, based on statin and PCSK9 inhibitor variants, with each disease using different outcome GWAS. As expected, the MR estimates show that genetically predicted lower LDL-cholesterol, based on statin or PCSK9 inhibitor genetic variants, reduced IHD, albeit with a slightly less marked effect for statin genetic variants on IHD in the UK Biobank. LDL-cholesterol lowering, based on statin or PCSK9 inhibitor genetic variants, was not associated with a lower risk of stroke. LDL-cholesterol lowering, based on statin genetic variants, had an association in a positive direction with AF while LDL-cholesterol lowering, based on PCSK9 inhibitor genetic variants, had an association in the opposite direction with AF. The contradictory results for stroke and AF compared to IHD could be due to differences in the underlying populations, but we replicated the findings for IHD and stroke in the UK Biobank. More parsimoniously, selection bias invalidating the

exclusion restriction assumption of instrumental variable analysis as shown in Figure 2b could be at play. Statins and PCSK9 inhibitors affect survival but death from IHD between randomization and recruitment precludes seeing their full effect on stroke and AF. Correspondingly, from the limited information about the underlying GWAS available, the AF cases appear to be older than the stroke cases who were older than the IHD cases (Supplementary Table 1).

Paradox explained

A previous MR study has similarly shown LDL-cholesterol lowering PCSK9 inhibitor genetic variants causally associated with IHD but not with stroke [45], and suggested the study showed the limits of MR [45]. Figure 2b suggests this anomalous finding for stroke may be the result of selection into the stroke GWAS dependent on surviving the harmful effects of LDL-cholesterol genetic variants and any other factors, such as blood pressure, obesity, diabetes and smoking, causing IHD and stroke without accounting for them. As such, the study in question does not show the limits to MR *per se* [45], but specifically an addressable violation of the instrumental variable assumption of exclusion-restriction in that MR study.

Insights obtained

Previous MR studies have suggested that smoking and hypertension protect against Alzheimer's disease (AD) [30]. On the one hand, these observations could reflect true causal relations. On the other hand, given smoking and hypertension adversely affect survival, these anomalous associations could be the result of shared etiology with other common conditions, such as IHD, that typically results in death before the average age of onset AD, and thereby preclude recruitment into the relevant studies. ApoE is a very well-established cause of AD [30]. ApoE also causing IHD would explain the anomalous inverse associations of smoking and hypertension with AD, as well as shedding new light on IHD, as shown in Figure 2c. More generally, selection bias from competing risk before recruitment in the presence of selective survival on exposure exists in population representative studies [46], because a population representative study is only the survivors of the underlying birth cohorts who formed the population. As such, MR phenome wide

association studies of the effect of one exposure on many outcomes may not be completely valid unless the exposure is non-fatal or consideration is given to accounting for shared etiology generating outcome specific competing risk.

Potential solutions

One potential solution is to conduct the MR analysis within a study where the genetic associations with the exposure are subject to the same biases as the genetic associations with the outcome. However, most exposures investigated, such as diet, are not subject to the same competing risk as typical disease outcomes. Moreover, many MR studies are conducted in two separate samples, likely subject to different biases. A conceptually simple way to address bias from surviving to recruitment on genetically predicted exposure in the presence of competing risk of the outcome is to use inverse probability weights of such survival, but these are often unknown. Alternatively, the path from instrument to outcome via survival could be blocked by adjusting for all factors underlying the shared etiology of outcome and competing risk. However, that these factors are not comprehensively known is often the motivation for an MR study. Moreover, such information is almost impossible to collect with sufficient precision to eliminate confounding by common and influential but hard to measure factors, such as lifestyle and socio-economic position. More generally, accounting for such factors is seen as a barrier to obtaining precise estimates in genetic studies [47], and an active area of research in MR studies [2], for which we have provided some clarification. However, it may be unclear exactly how to adjust as some common causes of the outcome and survival may also be consequences of the exposure thereby generating pleiotropic effects.

Few methods of sensitivity analysis exist to address selection bias in observational studies, and may be based on the assumption that the bias creates a false null [48] or a false association [49], rather than extending to bias generating an association in the reverse direction to the true effect. Nevertheless, at its simplest, in studies of diseases the level of selection bias due to surviving to recruitment on genetically

predicted exposure in the presence of competing risk of the outcome can be thought of as depending on the proportion of the exposed who are not available for recruitment because of prior death due to the genetically predicted exposure and the proportion of those who could have experienced the outcome who are not available for recruitment because of prior death from a competing risk. Assuming these proportions are independent and their corresponding probabilities do not sum to more than 1, then for an observed odds ratio greater than 1 the true odds ratio for genetically predicted exposure on disease can be estimated as the observed odds ratio multiplied by the ratio of the probability of surviving the exposure and the competing risk to the probability of surviving the exposure or the competing risk, as shown in the Appendix. Notably, the closer the true odds ratio is to the null the easier it is for a reversal to occur, as shown in Appendix Figure 1. For example, the odds ratio for the statin genetic variant (rs12916) on IHD is 0.96. Assuming, as suggested by RCTs, that statins have the same effect on IHD and stroke, it would only take 10% with the harmful statin genetic variant and 25% of potential stroke cases to have died from IHD or other competing risk before recruitment for the observed odds ratio to be exactly 1.0. If instead 40% of potential stroke cases had died from competing risk before recruitment then the odds ratio would reverse to 1.04. Notably, this selection bias may stem from the underlying genetic studies, which we have shown can be open to systematic selection bias from competing risk [50].

Checking the underlying genetic studies used in MR for validity directly is difficult because few genotypes are known to act via well-established physiological pathways with known effects, but other approaches are possible. First, replication using different genetic studies is increasingly possible. However replication studies could all be subject to the same biases [51]. Second, associations that change with recruitment age are indicative of a harmful effect and potential bias [52, 53], with the associations in younger people having the greatest validity [15] because selective survival on exposure and competing risk of the outcome prior to recruitment will usually be greater with age, as mortality rates for most conditions increase with age. However, this is a heuristic to identify flawed studies, rather than a solution; although

perhaps preferable to assuming a null MR estimate has more reliability than any other value [54]. Finally, the issue here of obtaining valid estimates in the presence of selective survival on exposure and competing risk of the outcome is similar to the issue of obtaining valid genetic estimates in other studies of survivors, i.e., patients. However, the current solution for obtaining valid estimates in genetic studies of patients relies on the assumption that the factors causing disease and disease progression differ [55].

Conclusion

Here, we have shown theoretically, empirically and analytically that MR studies can be open to quite severe selection bias arising from selective survival on genetically instrumented exposure when other causes of survival and outcome exist, i.e. competing risk before recruitment. Moreover, we have to some extent addressed the active area of research as to whether MR studies should adjust for other traits [2], and explained how this issue relates to the assumptions of instrumental variable analysis. Bias from such selection bias is likely to be least evident for MR studies of harmless exposures recruited shortly after genetic randomization with no competing risk, i.e., studies using birth cohorts with minimal attrition. Conversely, such bias is likely to be most evident for MR studies recruited at older ages examining the effect of a harmful exposure on an outcome subject to competing risk from shared etiology with other common conditions that occur earlier in life. Simple sensitivity analysis based on information about the exposure and the natural history of the outcome may help identify when such selection bias is likely. More methods of obtaining valid MR estimates, when the exclusion restriction is invalidated by selection bias stemming from competing risk, are required.

References

1. Taubes, G. *Researchers find a way to mimic clinical trials using genetics*. MIT Technology Review 2018; Available from: https://www.technologyreview.com/s/611713/researchers-find-way-to-mimic-clinical-trials-using-genetics/.
2. Davies, N.M., M.V. Holmes, and G. Davey Smith, *Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians.* Bmj, 2018. **362**: p. k601.
3. Holmes, M.V., M. Ala-Korpela, and G.D. Smith, *Mendelian randomization in cardiometabolic disease: challenges in evaluating causality.* Nat Rev Cardiol, 2017. **14**(10): p. 577-590.
4. Davey Smith, G. and S. Ebrahim, *What can mendelian randomisation tell us about modifiable behavioural and environmental exposures?* BMJ, 2005. **330**(7499): p. 1076-9.
5. Burgess, S., et al., *Use of Mendelian randomisation to assess potential benefit of clinical intervention.* BMJ, 2012. **345**: p. e7325.
6. Bareinboim, E. and J. Pearl, *Causal inference and the data-fusion problem.* Proc Natl Acad Sci U S A, 2016. **113**(27): p. 7345-52.
7. Greenland, S., *An introduction to instrumental variables for epidemiologists.* Int J Epidemiol, 2000. **29**(4): p. 722-9.
8. Maciejewski, M.L. and M.A. Brookhart, *Using Instrumental Variables to Address Bias From Unobserved Confounders.* JAMA, 2019.
9. Smith, G.D. and S. Ebrahim, *Mendelian randomization: prospects, potentials, and limitations.* Int J Epidemiol, 2004. **33**(1): p. 30-42.
10. Munafo, M. and G.D. Smith, *Biased Estimates in Mendelian Randomization Studies Conducted in Unrepresentative Samples.* JAMA Cardiol, 2018. **3**(2): p. 181.
11. Gkatzionis, A. and S. Burgess, *Contextualizing selection bias in Mendelian randomization: how bad is it likely to be?* Int J Epidemiol, 2018.
12. Hughes, R.A., et al., *Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis.* Epidemiology, 2019. **30**(3): p. 350-357.
13. Munafo, M.R., et al., *Collider scope: when selection bias can substantially influence observed associations.* Int J Epidemiol, 2017.
14. Hernan, M.A., et al., *Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses.* J Clin Epidemiol, 2016. **79**: p. 70-75.
15. Swanson, S.A., et al., *Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials.* Epidemiology, 2017. **28**(5): p. 653-659.
16. Vansteelandt, S., O. Dukes, and T. Martinussen, *Survivor bias in Mendelian randomization analysis.* Biostatistics, 2018. **19**(4): p. 426-443.
17. Nitsch, D., et al., *Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials.* Am J Epidemiol, 2006. **163**(5): p. 397-403.
18. Boef, A.G., S. le Cessie, and O.M. Dekkers, *Mendelian randomization studies in the elderly.* Epidemiology, 2015. **26**(2): p. e15-6.
19. Swanson, S.A., *A Practical Guide to Selection Bias in Instrumental Variable Analyses.* Epidemiology, 2019. **30**(3): p. 345-349.
20. Canan, C., C. Lesko, and B. Lau, *Instrumental Variable Analyses and Selection Bias.* Epidemiology, 2017. **28**(3): p. 396-398.
21. Smit, R.A.J., et al., *Survival Bias in Mendelian Randomization Studies: A Threat to Causal Inference.* Epidemiology, 2019. **30**(6): p. 813-816.
22. Kesteloot, H. and M. Decramer, *Age at death from different diseases: the flemish experience during the period 2000-2004.* Acta Clin Belg, 2008. **63**(4): p. 256-61.
23. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
24. Verbanck, M., et al., *Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases.* Nat Genet, 2018. **50**(5): p. 693-698.
25. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* Int J Epidemiol, 2017. **46**(6): p. 1985-1998.
26. Bowden, J., et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.
27. Pearl, J., *Causality: models, reasoning, and inference* 2nd ed. 2009, Cambridge: Cambridge University Press.
28. Hernandez-Diaz, S., E.F. Schisterman, and M.A. Hernan, *The birth weight "paradox" uncovered?* Am J Epidemiol, 2006. **164**(11): p. 1115-20.
29. Hernan, M.A., A. Alonso, and G. Logroscino, *Cigarette smoking and dementia: potential selection bias in the elderly.* Epidemiology, 2008. **19**(3): p. 448-50.

30. Kuzma, E., et al., *Which Risk Factors Causally Influence Dementia? A Systematic Review of Mendelian Randomization Studies.* J Alzheimers Dis, 2018. **64**(1): p. 181-193.
31. Riaz, H., et al., *Association Between Obesity and Cardiovascular Outcomes: A Systematic Review and Meta-analysis of Mendelian Randomization Studies.* JAMA Netw Open, 2018. **1**(7): p. e183788.
32. Mills, E.J., et al., *Efficacy and safety of statin treatment for cardiovascular disease: a network meta-analysis of 170,255 patients from 76 randomized trials.* Qjm, 2011. **104**(2): p. 109-24.
33. Chou, R., et al., *Statins for Prevention of Cardiovascular Disease in Adults: Evidence Report and Systematic Review for the US Preventive Services Task Force.* JAMA, 2016. **316**(19): p. 2008-2024.
34. Schmidt, A.F., et al., *PCSK9 monoclonal antibodies for the primary and secondary prevention of cardiovascular disease.* Cochrane Database Syst Rev, 2017. **4**: p. Cd011748.
35. Peng, H., et al., *The effect of statins on the recurrence rate of atrial fibrillation after catheter ablation: A meta-analysis.* Pacing Clin Electrophysiol, 2018.
36. Ettehad, D., et al., *Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis.* Lancet, 2016. **387**(10022): p. 957-67.
37. Emdin, C.A., et al., *Effect of antihypertensive agents on risk of atrial fibrillation: a meta-analysis of large-scale randomized trials.* Europace, 2015. **17**(5): p. 701-10.
38. Lu, Y., et al., *Efficacy and safety of long-term treatment with statins for coronary heart disease: A Bayesian network meta-analysis.* Atherosclerosis, 2016. **254**: p. 215-227.
39. Karatasakis, A., et al., *Effect of PCSK9 Inhibitors on Clinical Outcomes in Patients With Hypercholesterolemia: A Meta-Analysis of 35 Randomized Controlled Trials.* J Am Heart Assoc, 2017. **6**(12).
40. Ference, B.A., et al., *Mendelian Randomization Study of ACLY and Cardiovascular Disease.* N Engl J Med, 2019. **380**(11): p. 1033-1042.
41. Nikpay, M., et al., *A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease.* Nat Genet, 2015. **47**(10): p. 1121-30.
42. Malik, R., et al., *Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes.* Nat Genet, 2018.
43. Nielsen, J.B., et al., *Biobank-driven genomic discovery yields new insight into atrial fibrillation biology.* Nat Genet, 2018. **50**(9): p. 1234-1239.
44. Zhou, W., et al., *Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.* Nat Genet, 2018. **50**(9): p. 1335-1341.
45. Hopewell, J.C., et al., *Differential effects of PCSK9 variants on risk of coronary disease and ischaemic stroke.* Eur Heart J, 2018. **39**(5): p. 354-359.
46. Howe, C.J. and W.R. Robinson, *Survival-related Selection Bias in Studies of Racial Health Disparities: The Importance of the Target Population and Study Design.* Epidemiology, 2018. **29**(4): p. 521-524.
47. Pirinen, M., P. Donnelly, and C.C. Spencer, *Including known covariates can reduce power to detect genetic effects in case-control studies.* Nat Genet, 2012. **44**(8): p. 848-51.
48. Smith, L.H. and T.J. VanderWeele, *Bounding Bias Due to Selection.* Epidemiology, 2019. **30**(4): p. 509-516.
49. Thompson, C.A., Z.F. Zhang, and O.A. Arah, *Competing risk bias to explain the inverse relationship between smoking and malignant melanoma.* Eur J Epidemiol, 2013. **28**(7): p. 557-67.
50. Schooling, C.M., *Biases in GWAS: the dog that did not bark.* bioRxiv, 2019. https://www.biorxiv.org/content/10.1101/709063v1.
51. Munafo, M.R. and G. Davey Smith, *Robust research needs many lines of evidence.* Nature, 2018. **553**(7689): p. 399-401.
52. Thomas, S., et al., *The Milieu Interieur study - an integrative approach for study of human immunological variance.* Clin Immunol, 2015. **157**(2): p. 277-93.
53. Schooling, C.M., *Selection bias in population-representative studies? A commentary on Deaton and Cartwright.* Soc Sci Med, 2018. **210**: p. 70.
54. VanderWeele, T.J., et al., *Methodological challenges in mendelian randomization.* Epidemiology, 2014. **25**(3): p. 427-35.
55. Dudbridge, F., et al., *Adjustment for index event bias in genome-wide association studies of subsequent events.* Nat Commun, 2019. **10**(1): p. 1561.

Table 1: Effect of LDL-cholesterol lowering (1mmol/L) by genetically predicted statin and PCSK9 inhibitor [40] on IHD, stroke and AF using Mendelian randomization applied to the CARDIoGRAMplusC4D 1000 Genomes based GWAS of IHD [41], UK Biobank summary statistics from SAIGE for IHD and stroke [44], MEGASTROKE [42] for stroke and a study by Nielsen et al [43] for AF and the effects of statins and PCSK9 inhibitors on the same outcomes from meta-analysis of RCTs [33, 35, 38, 39]

		Estimates from							
		Mendelian Randomization of genetically predicted LDL lowering by				Meta-analysis of RCTs [33, 35, 38, 39] with lipid lowering by			
		Statins		PCSK9 inhibitors		Statins		PCSK9 inhibitors	
Disease	Source of genetic associations	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Ischemic heart disease	CARDIoGRAMplusC4D 1000 Genomes	0.59	0.44 to 0.81	0.55	0.35 to 0.87	0.69	0.61 to 0.77	0.72	0.64 to 0.81
	UK Biobank (SAIGE)	0.69	0.48 to 0.996	0.56	0.44 to 0.70				
All ischemic stroke	MEGASTROKE	1.01	0.72 to 1.41	1.08	0.97 to 1.22	0.71	0.62 to 0.82	0.80	0.67 to 0.96
	UK Biobank (SAIGE)	1.41	0.81 to 2.49	0.85	0.57 to 1.28				
Atrial fibrillation	Nielsen el al	1.14	0.92 to 1.42	0.85	0.71 to 1.01	0.47	0.30 to 0.75	na	

Supplementary table 1: Study details for the GWAS of IHD, stroke and AF

Study	Phenotype (Phewas code)	Cases	Non-cases	Mean age of cases	Phenotype definition	Adjusted for (non-genetic)
Cardiogram 1000 genomes GWAS [41]	Ischemic heart disease	60,801	123,504	n/a, possibly ~58 years	"Case status was defined by an inclusive CAD diagnosis (e.g. myocardial infarction (MI), acute coronary syndrome, chronic stable angina, or coronary stenosis >50%)"	Study-specific covariates (not age or sex)
UK biobank SAIGE [44]	CAD (411) Stroke (433) AF (427.2)	31,355 8,742 14,820	377,103 399,017 380,919	n/a	Phewas code based on self-report, hospital episodes and death	Sex, birth year, and principal components 1 to 4
MEGASTROKE [42]	All ischemic stroke	60,341	454,450	~69 years	Several different definitions used	Minimum of age and sex
AF [43]	Atrial fibrillation	60,620	970,216	~74 years	Usually based on ICD-9 427.3 and ICD-10 I48	Minimum of age (birth year) and sex

Figure 1: Directed acyclic graphs with instrument (Z), outcome (Y), exposure (X), confounders (U) and survival (S), where a box indicates selection, for a) a valid Mendelian randomization study, and b) a Mendelian randomization study with an invalid instrument through violation of the exclusion-restriction assumption via pleiotropy, c) a Mendelian randomization study with an invalid instrument through violation of the exclusion-restriction assumption via survival on instrument and shared etiology with the outcome (U_2), d) a Mendelian randomization study with an invalid instrument through violation of the exclusion restriction assumption via survival (S), competing risk of another disease (Y_2) and shared causes (U_2) with (Y_2) and the outcome (Y) and e) a Mendelian randomization illustrating both conditions which have to be met to satisfy the exclusion restriction assumption.

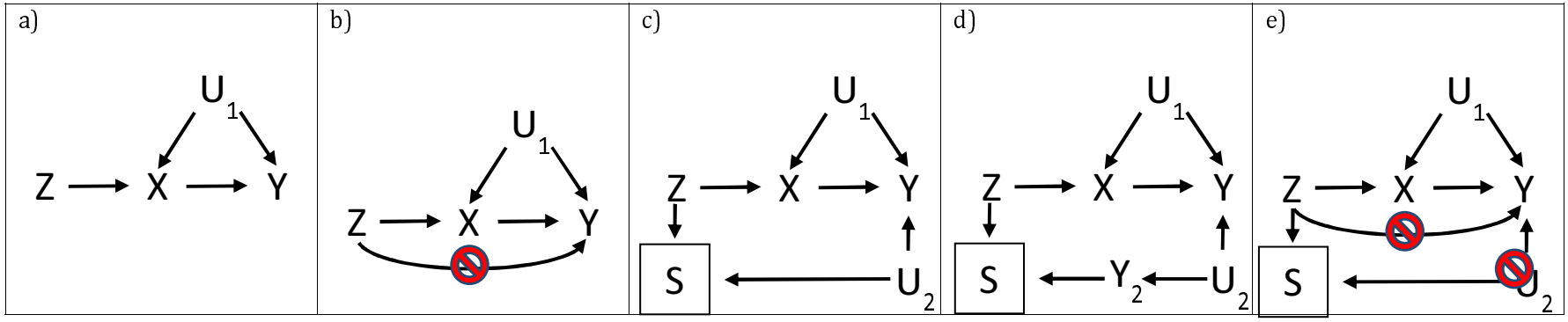
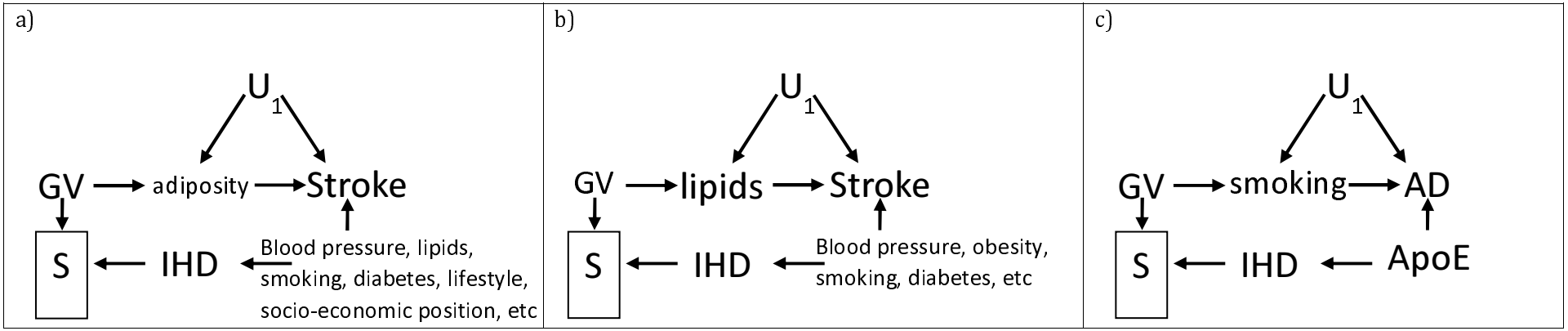


Figure 2: Directed acyclic graphs showing examples where selection bias could occur because of selection on survival (S), indicated by a box, on the instrument (GV) and on competing risk of ischemic heart disease (IHD) which shares causes with the outcome of interest, i.e., stroke or AD. U_1 indicates confounders of exposure on outcome. AD indicates Alzheimer's disease.



Appendix 1: A possible solution for recovering the causal effect in the presence of selection bias due to selecting on surviving the exposure and competing risk of the exposure in a case-control study

The fundamental issue of the selection bias in a case-control study is unknown information for the "missing" (or unselected) participants. Appendix Table 1 shows the possible mechanism generating a biased causal effect due to selection on surviving the exposure (E) and surviving competing risk (CR) of the outcome (D) in a case-control study.

Based on the observed data a', b', c', and d', the observed causal effect of E on D using an odds ratio (OR^{obs}) is,

$$OR^{obs} = \frac{a'/b'}{c'/d'} = \frac{a'd'}{b'c'}.$$

To obtain the true causal effect, we have to recover the data for the whole population, i.e. the birth cohorts who formed the population. Let P_E denote the proportion of participants unselected due to E, and let P_{CR} denote the proportion of participants unselected due to CR. Suppose P_E and P_{CR} are additive, and $0 < P_E + P_{CR} < 1$. We can construct the pattern of the unselected participants, as shown in Appendix Table 1. As such, the causal effect of E on D for the whole population can be estimated as follows,

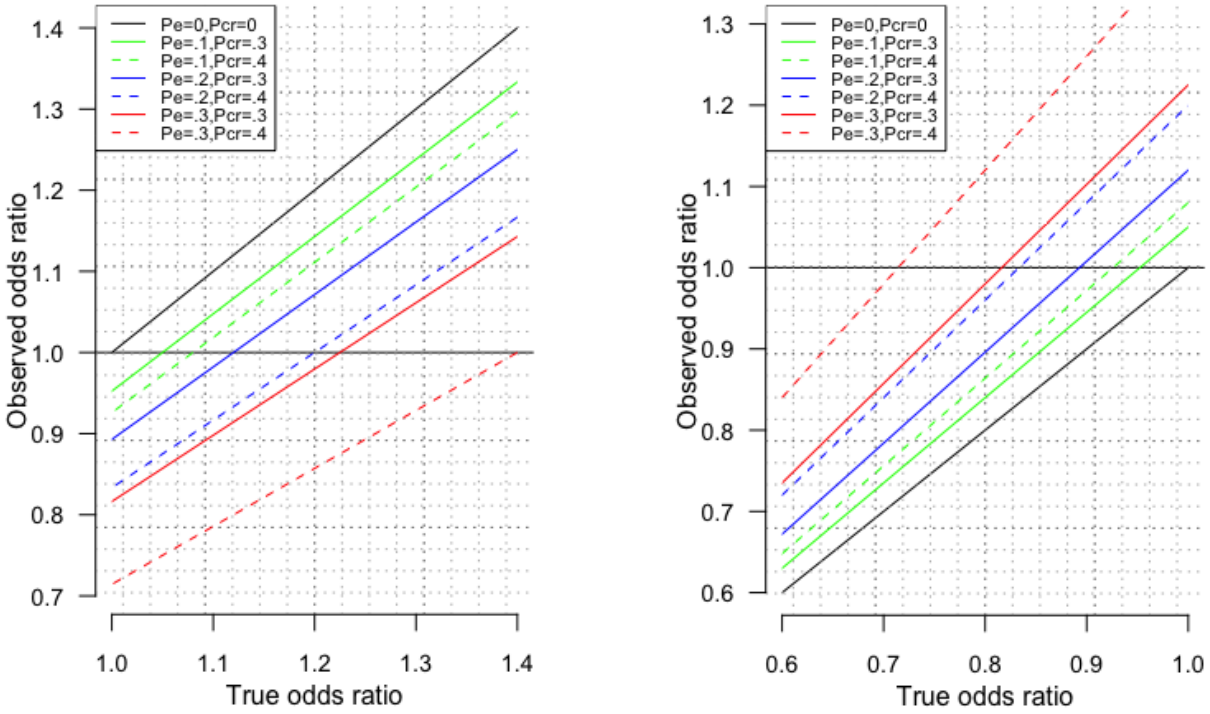
$$OR^{true} = \frac{ad}{bc} = \frac{\frac{a'/(1-P_E)}{b'/(1-P_E-P_{CR})}}{\frac{c'}{d'/(1-P_{CR})}} = OR^{obs} \times \frac{(1-P_E)(1-P_{CR})}{(1-P_E-P_{CR})}.$$

This relationship will be invalid if we replace the OR with a risk ratio.

Appendix Table 1. Possible mechanism for biased causal effects in a case-control study due to selection bias from surviving the exposure and competing risk of the outcome

		S = 1		S = 0	
		D = 1	D = 0	D = 1	D = 0
E = 1	CR = 1 CR = 0	a'	b'	P_E	$P_E + P_{CR}$
E = 0	CR = 1 CR = 0	c'	d'	0	P_{CR}
Observation		OR^{obs}			
Target		OR^{true}			

Note: S indexes selection status of participants, i.e., S = 1 indicates those selected and S = 0 indicates those unselected. D indexes outcome status, i.e., D = 1 indicates disease and D = 0 indicates no disease. E indexes exposure status, i.e., E = 1 indicates the exposed and E = 0 indicates unexposed. CR indicates competing risk (CR) of the outcome D; i.e., CR = 1 then D = 0, and if D = 1 and CR = 0. a', b', c', and d' are observed data about the selected participants.

Notably, the level of bias depends on the magnitude of the OR. A small OR, of the order of 1.05, as is typical in a genetic study, is much more vulnerable to a reversal of effect from selection bias due to selecting on surviving the exposure and surviving competing risk of the outcome than a larger OR, of the order of 1.50, as is typical in traditional observational studies. To clarify Appendix Figure 1 shows the observed OR plotted against the true OR for different combinations of selection on survival (P_E) and selection on competing risk of surviving the outcome (P_{CR})

Appendix Figure 1: Observed odds ratio against the True odds ratio in the presence of different proportions of death before recruitment due to the exposure (P_E) and different proportions of death before recruitment due to competing risk of the outcome (P_{CR}) for true odds ratios large than 1 (left hand side) and smaller than 1 (right hand side).