Selection bias when inferring the effect direction in Mendelian randomization

Sharon M. Lutz[1,2], **Kirsten Voorhies**[1], **Ann C. Wu**[1], **John Hokanson**[3], **Stijn Vansteelandt**[4,5], **Christoph Lange**[2]

[1]PRecisiOn Medicine Translational Research (PROMoTeR) Center, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[3]Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

[4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[5]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Keywords

Mendelian randomization; MR Steiger; selection bias

Dear Editor,

In our recent paper (Lutz, Wu, et al., 2021), we examined the ability of the MR Steiger approach by Hemani et al. (2017) to infer the effect direction under pleiotropy, measurement error, and unmeasured confounding. In their letter to the editor, Hemani et al focus on two aspects of our paper: the data analysis and the role of unmeasured confounding. Hemani et al also describe additional problems/issues with the original MR Steiger method (Hemani et al., 2017) that we did not discuss in our original paper.

In our paper (Lutz, Wu, et al., 2021), the purpose of our data analysis was to provide an example of when the MR Steiger approach can provide obviously incorrect conclusions. We used the MR Steiger approach to examine the role of current smoking status (current vs. former smoker) on forced expiratory volume in 1 s (FEV_1) in the COPDGene study, a case-control study of chronic obstructive pulmonary disease (COPD) in current and former smokers. In our analysis, the MR Steiger approach concluded that FEV_1 causes current

Correspondence: Sharon M. Lutz, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Dr, Suite 401 East, Boston, MA 02215, USA., smlutz@hsph.harvard.edu.

smoking status. We attributed this result to possible pleiotropy between the chromosome 15q25 region with smoking behavior and lung function. We agree with Hemani et al that this result could also have been due to selection bias, measurement error, unmeasured confounder, or other reasons that can also bias the MR Steiger approach. In addition, Hemani et al. raise the very important point that the MR Steiger approach is susceptible to selection bias and can provide spurious results in ascertained studies, possibly severely limiting its application to case/control studies.

Since the COPDGene study recruited COPD cases and controls who smoked at least 10 pack-years of cigarettes, Hemani et al. reperformed the analysis in the UK Biobank. They consider FEV_1 in the UK Biobank ($n = 421,986$) and four smoking phenotypes: ever versus never-smoking status ($n = 461,066$) in the UK Biobank, ever stopped smoking for 6+ months ($n = 113,230$) in the UK Biobank, cigarettes per day from a meta-analysis (Liu et al., 2019) ($n = 337,334$), and current vs former smoking status from a meta-analysis by Furberg et al. (Tobacco and Genetics Consortium, 2010) ($n = 41,969$). We do not understand why Hermani et al. did not consider current versus former smoking status in the UK Biobank for their reanalysis, although it is available in the UK Biobank and is the phenotype that we used in our original analysis. For the separate SNP analysis, Hemani et al only considered current vs former smoking status. However, rather than using the unascertained UK Biobank data for this analysis, Hermani et al. used the meta-analysis by Furberg et al. (Tobacco and Genetics Consortium, 2010), which is based on multiple studies including case/control studies. Unless the goal was to evaluate selection bias, the more appropriate data for this analysis would have been the UK Biobank.

For our reanalysis in the UK Biobank, we used the code provided by Hemani et al. to perform the separate SNP analysis for the other three smoking phenotypes and the UK Biobank's current versus former smoking status. All the code we used to perform this analysis is included in the supplement. As seen in Table 1 for both smoking phenotypes (ever versus never-smoking status and current versus former smoking status in the UK Biobank), there is a significant p-value for the Steiger correlation, and the incorrect direction is detected for some of the SNPs, that is, FEV_1 causes smoking status. This differs from the results by Hemani et al., which used current versus former smoking status from a meta-analysis and not the UK Biobank.

Regarding Hemani's criticism about our choice for a single SNP analysis, we note that the original paper established the MR Steiger method (Hemani et al., 2017) and its features (simulation studies) for the single SNP analysis. However, Hemani et al. (2017) suggest that the MR Steiger directionality test should be performed for multiple SNPs. The MR Steiger (Hemani et al., 2017) paper states: "We also note that it is straightforward to extend the MR Steiger approach to multiple instruments, requiring only that the total variance explained by all instruments be calculated under the assumption that they are independent." Nevertheless, the extension to multiple SNPs is not provided in the paper. In the TwoSampleMR R-package, Hemani et al implement a multiple SNP approach by calculating the correlation between the phenotype (i.e., exposure or outcome) and the SNPs as the square root of the sum of the correlations for each SNP and phenotype pair. This approach to combine the correlations is not explicitly discussed at all in the original paper

and it is not clear why this approach is even valid, as the proposed combining of correlations is not range consistent. For that reason, we considered the SNPs separately in our analysis.

Regarding their point about unmeasured confounding, we agree with Hemani et al. that the MR Steiger method can be biased in the presence of unmeasured confounding. This is an error on our side. Our DAG reasoning did not alert us to this subtle point: it shows that unmeasured confounding does not bias the correlation when it is zero, but is silent as to how a nonzero correlation may be affected. Unmeasured confounding indeed affects phenotypic variability, which in turn may distort comparisons of the magnitude of correlations, as made by the MR Steiger approach (Lutz, Voorhies, et al., 2021).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

Hemani G, Tilling K, & Davey Smith G (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genetics, 13, e1007081. [PubMed: 29149188]

Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, Datta G, Davila-Velderrain J, McGuire D, Tian C, Zhan X, Research T, HUNT All-In P, Choquet H, Docherty AR, Faul JD, Foerster JR, Fritsche LG, Gabrielsen ME, … Vrieze S (2019). Association studies of up to 1.2 million individuals yield new insights in the genetic etiology of tobacco and alcohol use. Nature Genetics, 51(2), 237–244. [PubMed: 30643251]

Lutz SM, Voorhies K, Wu AC, Hokanson J, Vansteelandt S, & Lange C (2022). The influence of unmeasured confounding on the MR Steiger approach. Genetic Epidemiology, 46(2), 139–141. 10.1002/gepi.22442 [PubMed: 35170805]

Lutz SM, Wu AC, Hokanson JE, Vansteelandt S, & Lange C (2021). Caution against examining the role of reverse causality in Mendelian randomization. Genetic Epidemiology, 45, 445–454. [PubMed: 34008876]

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature Genetics, 42(5), 441–447. [PubMed: 20418890]

Author Manuscript   Author Manuscript   Author Manuscript   Author Manuscript

TABLE 1

We considered the single SNP analysis for FEV_1 from the UK Biobank and the following smoking phenotypes

SNP	Chr	Ever/Never smoker (UK Biobank)		Quit smoking 6+ months (UK Biobank)		Cigarettes per day (Lui et al., 2019)		Former/current (UK Biobank)			
		Correct Steiger	Steiger p value	Correct Steiger	Steiger p value	Correct Steiger	Steiger p value	Cor (G,X)*	Cor (G,Y)	Correct Steiger	Steiger p value
rs56113850	19	FALSE	0.10	FALSE	0.72	TRUE	1.2e–33	−0.024	−0.01	TRUE	9e–5
rs7260329	19	FALSE	0.40	FALSE	0.82	TRUE	4.9e–6	0.009	0.01	FALSE	0.80
rs11858836	15	**FALSE**	0.02	TRUE	0.94	TRUE	1.5e–64	0.004	−0.02	**FALSE**	2e–11
rs72738786	15	**FALSE**	8.0e–3	FALSE	0.99	TRUE	2.7e–82	0.004	−0.02	**FALSE**	2e–11
rs11633958	15	**FALSE**	0.02	TRUE	0.91	TRUE	1.4e–81	0.005	−0.02	**FALSE**	3e–12
rs8192482	15	**FALSE**	0.03	TRUE	0.86	TRUE	1.7e–82	0.005	−0.02	**FALSE**	3e–12
rs2869548	15	**FALSE**	0.04	TRUE	0.77	TRUE	3.6e–73	0.004	−0.02	**FALSE**	2e–11

Note: The "Correct Steiger" column is TRUE if smoking causes FEV_1 and FALSE otherwise. Results in bold indicate that the correct Steiger direction was FALSE and the $p < 0.05$. Note that Cor(G,X)*, the correlation between the SNP G and former/current smoking status in the UK Biobank, is calculated using the recommended get_r_from_lor function from the TwoSampleMR package since the exposure is binary. Note that the SNPs are in LD and caution should be used for the overall test.