

1 **Exploring and mitigating potential bias when genetic instrumental variables are associated with**  
2 **multiple non-exposure traits in Mendelian randomization**

3 Qian Yang,<sup>1,2\*</sup> Eleanor Sanderson,<sup>1,2</sup> Kate Tilling,<sup>1,2,3</sup> M Carolina Borges,<sup>1,2†</sup> Deborah A Lawlor<sup>1,2,3†</sup>

4 <sup>1</sup>MRC Integrative Epidemiology Unit at the University of Bristol

5 <sup>2</sup>Population Health Sciences, Bristol Medical School, University of Bristol

6 <sup>3</sup>National Institute for Health Research Bristol Biomedical Center, University Hospitals Bristol NHS  
7 Foundation Trust and University of Bristol

8

9 \*Corresponding author

10 Address: Office room OF28, Oakfield House, Oakfield Grove, Clifton, Bristol, BS8 2BN, UK

11 Email: [qian.yang@bristol.ac.uk](mailto:qian.yang@bristol.ac.uk)

12 †Deborah A Lawlor and M Carolina Borges made equal contributions and are joint senior authors.

13

14 **Key words:** Mendelian randomization, population stratification, pleiotropy, causal diagram, UK

15 Biobank

16 **Abstract**

17 **Background:** Our aim is to produce guidance on exploring and mitigating possible bias when genetic  
18 instrumental variables (IVs) associate with traits other than the exposure of interest in Mendelian  
19 randomization (MR) studies.

20 **Methods:** We use causal diagrams to illustrate scenarios that could result in IVs being related to  
21 (non-exposure) traits. We recommend that MR studies explore possible IV-non-exposure  
22 associations across a much wider range of traits than is usually the case. Where associations are  
23 found, confounding by population stratification should be assessed through adjusting for relevant  
24 population structure variables. To distinguish vertical from horizontal pleiotropy we suggest using  
25 bidirectional MR between the exposure and non-exposure traits and MR of the effect of the non-  
26 exposure traits on the outcome of interest. If vertical pleiotropy is plausible, standard MR methods  
27 should be unbiased. If horizontal pleiotropy is plausible, we recommend using multivariable MR to  
28 control for observed pleiotropic traits and conducting sensitivity analyses which do not require prior  
29 knowledge of specific invalid IVs or pleiotropic paths.

30 **Results:** We applied our recommendations to an illustrative example of the effect of maternal  
31 insomnia on offspring birthweight in the UK Biobank. We found little evidence that unexpected IV-  
32 non-exposure associations were driven by population stratification. Three out of six observed non-  
33 exposure traits plausibly reflected horizontal pleiotropy. Multivariable MR and sensitivity analyses  
34 suggested an inverse association of insomnia with birthweight, but effects were imprecisely  
35 estimated in some of these analyses.

36 **Conclusions:** We provide guidance for MR studies where genetic IVs associate with non-exposure  
37 traits.

38 (word limit: 250; word count: 247)

## 39 **Key messages**

- 40 • Genetic variants are increasingly found to associate with more than one social, behavioural or  
41 biological trait at genome-wide significance, which is a challenge in Mendelian randomization  
42 (MR) studies.
- 43 • Four broad scenarios (i.e. population stratification, vertical pleiotropy, horizontal pleiotropy and  
44 reverse causality) could result in an IV-non-exposure trait association.
- 45 • Population stratification can be assessed through adjusting for population structure with  
46 individual data, while two-sample MR studies should check whether the original genome-wide  
47 association studies have used robust methods to properly account for it.
- 48 • We apply currently available MR methods for discriminating between vertical and horizontal  
49 pleiotropy and mitigating against horizontal pleiotropy to an example exploring the effect of  
50 maternal insomnia on offspring birthweight.
- 51 • Our study highlights the pros and cons of relying more on sensitivity analyses without  
52 considering particular pleiotropic paths versus systematically exploring and controlling for  
53 potential pleiotropic paths via known characteristics.

## 54 Introduction

55 Mendelian randomization (MR) is a special case of instrumental variable (IV) analysis where single  
56 nucleotide polymorphisms (SNPs) randomly allocated at conception are used as the IVs.<sup>(1, 2)</sup> MR  
57 requires three key assumptions: first, IVs are strongly associated with an exposure of interest  
58 (relevance); second, there are no common causes between IVs and an outcome of interest  
59 (independence); and third, IVs influence the outcome only through the exposure (exclusion  
60 restriction).<sup>(1, 2)</sup> While the relevance assumption can be tested, the independence and exclusion  
61 restriction assumptions are difficult to verify and only their plausibility can be explored.<sup>(3)</sup> One  
62 common approach to date is to test for associations between genetic IVs and a range of non-  
63 exposure traits in either one- or two-sample setting as a way to assess the specificity of the genetic  
64 IV.<sup>(4-7)</sup>

65 With increasing sizes of genome-wide association studies (GWAS), and more extensive coverage of  
66 the genome due to imputation with more comprehensive panels, SNPs are increasingly found to  
67 associate with multiple traits.<sup>(8, 9)</sup> Therefore, we aim to develop guidance for assessing potential  
68 violations of independence and exclusion restriction assumptions when genetic IVs are associated  
69 with other (non-exposure) traits. This paper is laid out as follows. In section 1, we use directed  
70 acyclic graphs (DAGs) to illustrate four scenarios that could result in an association of a genetic IV  
71 with a non-exposure trait and highlight which scenarios would bias MR estimates. In section 2, we  
72 describe different methods for discriminating between scenarios and methods for mitigating against  
73 potential bias for both one- and two-sample settings. In section 3, we apply this framework to an MR  
74 analysis exploring the potential causal relationship between maternal insomnia and offspring  
75 birthweight in the UK Biobank (UKB). In section 4, we end with a discussion of our  
76 recommendations.

## 77 **Scenarios that could explain associations of genetic IVs with multiple traits**

78 There are four broad scenarios consistent with genetic IVs (Z) being associated with multiple traits  
79 (Table 1). **Population stratification** (PS; DAGs 1.1-1.3 in Table 1), might occur due to the study  
80 including subgroups of people with different ancestry or who were born or live in different  
81 geographical locations. If the distribution of SNPs and of non-exposure traits (W) differs by these  
82 subgroups, then this PS is a common cause of Z and W and generates an association between  
83 them.<sup>(10)</sup> As an example of this, evidence shows that place of birth in UKB has been associated with  
84 genetic IVs for education, height and body mass index (BMI), and also with health outcomes.<sup>(11)</sup> If PS  
85 affects the distribution of Z and the outcome (Y) directly or via W, PS could confound MR estimates  
86 (DAGs 1.1-1.2). This would represent a violation of the independence assumption. If PS confounds  
87 the Z-exposure (X) association, Z could still be used to estimate the unbiased effect of X on Y if PS did  
88 not affect Y independently of X (DAG 1.3).

89 Pleiotropy refers to the association of a SNP with multiple phenotypes, and has two types: vertical  
90 (also known as spurious or false) and horizontal (also known as genuine or true).<sup>(12)</sup> In the scenario of  
91 **vertical pleiotropy** (DAGs 2.1-2.3 in Table 1), Z is a cause of X, which in turn affects Y. Despite  
92 pleiotropic associations of Z with X and W, the effect of Z on Y is fully mediated by X. Therefore, the  
93 exclusion restriction assumption is not violated.<sup>(13)</sup> In the scenario of **horizontal pleiotropy** (DAGs  
94 3.1-3.3 in Table 1), Z is a cause of X and W, and both of them affect Y independently. This violates  
95 the exclusion restriction assumption leading to potential bias in MR estimates.<sup>(13)</sup>

96 In the scenario of **reverse causality** (DAG 4.1 in Table 1), Z is really a primary cause of Y, which in  
97 turn affects X. As such, inclusion of Z into IVs for X would give a biased X-Y association due to a  
98 violation of the exclusion restriction assumption.<sup>(14)</sup> With respect to the focus of this paper, this  
99 would only result in an association of Z with W if Z directly or indirectly influences W (DAG 4.1). We  
100 have included this scenario for completeness. However, exploration of Z-W associations is not a  
101 good way of identifying causal directions between X and Y. Bidirectional MR and Steiger

- 102 directionality test should be more suitable for exploring causal directions between any two traits,<sup>(14)</sup>
- 103 and will be described in recommendation 4 of the next section.

104 **Recommendations for exploring above scenarios and to obtain unbiased MR estimates**

105 Having described the different scenarios that could result in genetic IVs relating to non-exposure  
106 traits below we provide a list of recommendations for first identifying such non-exposure traits and  
107 then exploring which are likely to bias the main MR results and how that might be mitigated  
108 (summarised in Table 1).

109 **1. Searching more thoroughly for genetic IV-non-exposure trait associations**

110 To date most MR studies have explored associations of genetic IVs with potential confounders of  
111 exposure-outcome associations. By definition exposure-outcome confounders are unlikely to  
112 influence genetic variants (which are fixed at conception) and the association of genetic IVs with  
113 several non-exposure traits are likely to reflect violation of the exclusion restriction via pleiotropy  
114 (i.e. the direction of the arrow will go from genetic variants to the confounders rather than the other  
115 way). Consequently, selection of non-exposure trait associations should aim to explore violation of  
116 the independence and exclusion restriction assumptions and their specific mechanisms by exploring  
117 associations with any risk factors for the outcome, rather than focus solely potential exposure-  
118 outcome cofounders.

119 Once this is acknowledged there are two broad approaches that could be used to identify non-  
120 exposure traits that genetic IVs might influence in either one- or two-sample MR. One is to use  
121 prior/existing knowledge of the key causes of the outcome and then examine whether the genetic  
122 exposure-IV relates to any of these non-exposure causes of the outcome. The second is to undertake  
123 a hypothesis free comprehensive genotype-to-phenotype (also known as Phenome-wide) approach,  
124 in which we use automated systems to explore all possible associations of our genetic IVs.<sup>(15, 16)</sup>  
125 There are differing pros and cons of these two approaches including different challenges relating to  
126 balance between valid application of the following recommendations 3-5 versus a greater reliance  
127 on sensitivity analyses in recommendation 6. We explore this further in the discussion.

128 **2. Assessing the impact of population stratification**

129 In one-sample MR with individual participant data, we recommend exploring associations of Z with  
130 as many indicators of PS as possible. These could include place (country, region, town,  
131 longitude/latitude) of birth and residence, study centre, and genetic principal components of  
132 ancestral background. Adjusting the Z-W associations for these sources of PS and exploring whether  
133 this alters the association is also useful for exploring PS. If the association was attenuated after  
134 adjustment, it would suggest that the Z-W association may be driven by PS. In two-sample MR using  
135 summary statistics, the data have typically been generated a priori and thus the investigators are  
136 limited in what they can do to account for PS. However, they can still check whether the original  
137 GWAS has used robust genomic control methods to properly account for PS. Newly developed two-  
138 sample MR methods (MR-PRESSO,<sup>(17)</sup> GSMR,<sup>(18)</sup> LCV<sup>(19)</sup> and GIV<sup>(20)</sup>) may not be able to overcome PS  
139 either if PS is not controlled in the original GWAS, as acknowledged by Koellinger *et al.*<sup>(10)</sup>

140 **3. Assessing bias due to horizontal pleiotropy by using MR to explore the W-Y association**

141 After excluding the possibility of bias by population stratification, it is important to investigate  
142 whether unexpected Z-W associations might be explained by horizontal pleiotropy. We recommend  
143 first undertaking MR to explore whether there is evidence for the effect of W on Y. This requires  
144 valid genetic IVs for W, which may not always be available, and sufficient statistical power to  
145 precisely estimate the W-Y association. It is also important to consider the strength of the genetic  
146 IVs for W, as weak instrument bias would tend to bias the estimate towards the observational  
147 association in one-sample MR but to the null in two-sample MR with non-overlapping samples and  
148 increase the standard errors of the estimate.<sup>(21)</sup> If there are valid and strong genetic IVs for W and  
149 these provide (convincing) evidence that W does not affect Y, then there cannot be violation of the  
150 exclusion restriction criteria via W. If there is evidence for an effect of W on Y, or it is not possible to  
151 determine this, then bidirectional MR of the effect of X-W versus W-X is valuable (next point).



152 **4. Distinguishing vertical from horizontal pleiotropy by testing causal directions between X and**

153 **W**

154 If there is no causal effect between X and W, horizontal pleiotropy (DAG 3.1) would be more  
155 plausible than vertical pleiotropy (DAGs 2.1-2.3). If bidirectional MR suggests that X causes W,  
156 vertical pleiotropy (DAGs 2.1-2.2) may be more plausible than horizontal pleiotropy (DAGs 3.1-3.2),  
157 although we could not fully rule out the possibility that W mediates the effect from Z to Y partly  
158 independently of X (DAG 3.3). However, if bidirectional MR suggests that W causes X, W could be a  
159 confounder of X and Y and horizontal pleiotropy (DAG 3.2) may be more plausible than vertical  
160 pleiotropy (DAGs 2.1-2.2), although we could not fully rule out the possibility that W cannot affect Y  
161 independently of X (DAG 2.3).

162 Bidirectional MR can be conducted in either one- or two-sample setting,<sup>(22, 23)</sup> but could be  
163 misleading when there is marked difference in statistical power between X-W versus W-X  
164 associations. For example, if the power for W-X association is low (relative to the power for X-W  
165 association) it may appear that there is no causal effect of W on X even in the presence of such an  
166 effect. Additionally, overlapping SNPs in the GWAS of X and W can make it unclear which SNPs to  
167 select as valid IVs for X and W in bidirectional MR.<sup>(24)</sup> In two-sample setting, Steiger directionality test  
168 can help to identify (independent) valid IVs for X or W by comparing the variance explained by each  
169 SNP in X to that in W, as it assumes that a valid IV should explain more variance in trait A than in trait  
170 B if trait B is a downstream effect of the trait A.<sup>(25)</sup> However, Steiger directionality test could be  
171 misleading if measurement errors in X and W are substantially different.<sup>(25)</sup> For example, insomnia is  
172 measured much less accurately than height or BMI in UKB (see real data example in the next  
173 section).

174 **5. Adjusting for potential horizontal pleiotropic effects via known non-exposure traits**

175 Where there is evidence (from 3 and 4 above) that there may be bias due to horizontal pleiotropy  
176 (DAGs 3.1-3.2) from specific non-exposure traits (W), multivariable Mendelian randomization

177 (MVMR) can be used to obtain unbiased estimates in one- and two-sample settings.<sup>(26)</sup> MVMR  
178 requires not only  $IV_X$  (IV for X)-X and  $IV_W$ -W associations but also  $IV_X$ -W and  $IV_W$ -X associations, which  
179 means two-sample MR studies using summary statistics have to access to full results of the original  
180 GWAS. If W mediates both Z-Y and X-Y associations (DAG 3.3), controlling for W in MVMR obtains  
181 the direct effect rather than the total effect of X on Y, while its total effect should be estimated by  
182 using a subset of SNPs only related to X.<sup>(26)</sup> MVMR can also be used to estimate direct effects of  
183 correlated traits on an outcome as long as the genetic IVs independently predict each trait.  
184 Limitations of MVMR have been discussed by Sanderson *et al.*, e.g. the strengths of IVs may  
185 decrease dramatically after attempting to including many non-exposure traits in the estimation.<sup>(26)</sup>

## 186 **6. Exploring and controlling for bias due to horizontal pleiotropy via unknown traits**

187 It is possible that MVMR adjusting for horizontal pleiotropy via known/measured traits is still biased  
188 by unknown/unmeasured traits. Therefore, sensitivity analyses that explore bias due to unbalanced  
189 ‘unmeasured’ horizontal pleiotropy will still be required. In one- and two-sample MR, we  
190 recommend initial exploration of this by assessing between SNP heterogeneity.<sup>(27)</sup> This should be  
191 done even if SNPs are being combined into a single polygenic risk score (PRS). In one-sample MR  
192 heterogeneity is commonly explored by ‘overidentifying’ tests,<sup>(28)</sup> while in two-sample MR using  
193 summary data the Cochran’s Q statistic is an equivalent test.<sup>(27)</sup> If the exposure causes the outcome  
194 and IVs are valid, we expect the effect of the IV on the outcome to be proportional to its effect on  
195 the exposure across genetic IVs. Therefore, heterogeneous causal estimates across genetic IVs are  
196 indicative of invalid IVs. Most of the sensitivity analyses that have been developed for addressing  
197 horizontal pleiotropy aim at exploring the presence or being robust to heterogeneous (potentially  
198 invalid) IVs. Table 2 summarises the commonly used methods (i.e. sisVIVE<sup>(29)</sup> for one-sample MR and  
199 MR-Egger,<sup>(30)</sup> weighted median,<sup>(31)</sup> weighted mode,<sup>(32)</sup> MR-PRESSO,<sup>(17)</sup> MR-TRYX<sup>(33)</sup> for two-sample  
200 MR), including their additional assumptions. It is important to recognise that (i) heterogeneity tests

- 201 can only be used where there are multiple SNPs, (ii) some methods are statistically inefficient and
- 202 (iii) most methods have been developed for the two-sample MR setting.

203 **Real data example**

204 We use the effect of maternal insomnia on offspring birthweight as a motivating example, as it has  
205 been suggested that having insomnia and other forms of sleep disturbance may be associated with  
206 lower offspring birthweight though results are inconsistent.<sup>(24, 34, 35)</sup> We explore this question in UK  
207 Biobank women,<sup>(36)</sup> using a PRS that combines 80 genome-wide significant SNPs (listed in  
208 Supplementary Data 1) from the largest GWAS of insomnia in women.<sup>(37)</sup> We tested associations of  
209 the PRS with six observed traits (maternal height, BMI, age at first live birth, education, frequency of  
210 alcohol intake and ever smoking) that are known to (or could plausibly) influence offspring  
211 birthweight, and found that the PRS was associated with all of them (Figure 1). We demonstrate how  
212 to use the above recommendations in both one- and two-sample MR analyses, with full details in  
213 Supplementary Methods.

214 UKB is a cohort of 503,325 men and women who were on the National Health Service registry, aged  
215 between 40-69 years and living within 25 miles from one of 22 assessment centres.<sup>(36)</sup> One-sample  
216 MR included genetically unrelated women of European descent who reported frequency of  
217 insomnia, had experienced at least one live birth and reported the birthweight of their first live born  
218 child (N=165,184). Supplementary Table 1 summarises how each variable used here were measured  
219 in UKB and coded in our example. We also randomly split those genetically unrelated women of  
220 European descent into two groups (Supplementary Figure 1) to obtain SNP specific summary  
221 statistics for two-sample MR in this illustrative example. We selected the SNPs for these analyses  
222 from the published GWAS of insomnia<sup>(37)</sup>, height,<sup>(38)</sup> BMI,<sup>(39)</sup> age at first live birth<sup>(40)</sup> and education<sup>(41)</sup>  
223 in women and from the previous GWAS of frequency of alcohol intake and ever smoking in UKB men  
224 and women.<sup>(42)</sup> We obtained both SNP-exposure and SNP-outcome results from both of the random  
225 sub-samples and the pooled results from analyses in which sample 1 was used for SNP-exposure and  
226 sample 2 for SNP-outcome with those in which sample 1 was used for SNP-outcome and sample 2  
227 for SNP-exposure (Supplementary Methods).

## 228 ***Exploring the role of population stratification***

229 Each additional allele in the insomnia PRS was associated with a variation of -0.004 (95% confidence  
230 interval [CI]: -0.007, -0.001) year in age at recruitment,  $-3.7 \times 10^{-7}$  (95% CI:  $-6.7 \times 10^{-7}$ ,  $-6.4 \times 10^{-8}$ ) metre  
231 (M) in longitude of birthplace and  $2.1 \times 10^{-7}$  (95% CI:  $5.7 \times 10^{-8}$ ,  $3.5 \times 10^{-7}$ ) M in latitude of birthplace.  
232 There was evidence of some variation in the mean PRS across 22 UKB assessment centres  
233 (Supplementary Figure 2; P-value =  $9.2 \times 10^{-8}$ ). After adjusting for genetic array, top 40 genetic  
234 principal components, participants' age, birthplace and assessment centre, associations of the PRS  
235 with height, BMI, education, frequency of alcohol consumption and ever smoking were not  
236 attenuated (Figure 1) suggesting these associations are unlikely to be driven by PS. The association  
237 of the insomnia PRS with age at first live birth was slightly attenuated to the null (Figure 1),  
238 suggesting some of this may be due to confounding by PS. However, we obtained similar estimates  
239 in the MR analyses before and after adjustment for sources of PS (Supplementary Table 2).

## 240 ***Distinguishing vertical from horizontal pleiotropy and accounting for horizontal pleiotropy***

241 We searched for GWAS to identify genetic IVs for each of the six non-exposure traits that were  
242 conducted in samples independent of UKB and had results presented in women only. However, we  
243 were unable to find such GWAS of frequency of alcohol intake or ever smoking. Full details of the  
244 selected SNPs are provided in Supplementary Data 1. We found that height, BMI and frequency of  
245 alcohol consumption were more likely to reflect vertical pleiotropy or not associated with  
246 birthweight (Figure 2), suggesting the associations of the PRS with them were unlikely to introduce  
247 bias. These findings have some consistency with previous MR studies.<sup>(18, 43-45)</sup> However, age at first  
248 live birth, education and ever smoking were plausible sources of horizontal pleiotropy (Figure 2).  
249 After adjusting for these in MVMR, the effect estimates of insomnia on birthweight attenuated  
250 towards the null compared to univariable MR (Figure 3), though results are imprecise.  
251 In sensitivity analysis in one-sample MR, sisVIVE (full results in Supplementary Data 2) suggested that  
252 the association of insomnia with birthweight was greater than seen with univariable TSLS (-87 [95%

253 CI: -182, 7] grams). In the two sample MR results from all sensitivity analyses were directionally  
254 consistent with the main IVW estimate, though for several the CIs were very wide; IVW, MR-PRESSO  
255 and MR-TRYX supported an inverse association of maternal insomnia with offspring birthweight with  
256 CIs that did not include the null (Figure 4). The MR-Egger intercept suggested little evidence of  
257 unbalanced horizontal pleiotropy (p-value = 0.732 for dataset A on B and 0.763 for B on A; full  
258 results in Supplementary Figure 3). Whilst between SNP heterogeneity was less when MR-TRYX was  
259 used (in comparison to the IVW analyses) the point estimates were very similar between it and IVW  
260 (Figure 4 and Supplementary Figure 4).

### 261 ***Identifying more potential sources of violation of MR assumptions using a phenome-wide*** 262 ***approach***

263 In this motivating example we only explored the six traits that we had *a priori* selected for checking,  
264 based on prior knowledge that these were key risk factors for variation in birthweight. However,  
265 there may be value in exploring a wider range of potential violating paths (recommendation 1).  
266 Therefore, we undertook a comprehensive search for previously identified genome-wide significant  
267 associations of the 80 SNPs in the insomnia PRS using Phenoscanner.<sup>(15)</sup> This amounted to 478  
268 associations that included 42 SNPs, among which 34 SNPs were associated with at least one trait  
269 apart from sleep (full results in Supplementary Data 3). We did not examine these further but  
270 discuss in the discussion section the pros and cons of different approaches to identifying genetic IVs  
271 associations with non-exposure traits.

### 272 ***Exploring reverse causality***

273 Whilst temporally it is hard to conceive of birthweight influencing maternal insomnia, birthweight is  
274 a proxy for fetal growth, which could influence maternal insomnia symptoms. To explore this  
275 possibility, we would require offspring (fetal) genetic variants that are robustly related to their fetal  
276 growth. Whilst there are no GWAS currently of maternal or fetal contribution to fetal growth (e.g.  
277 assessed by repeat ultrasound scan) there are GWAS of own (i.e. fetal) genetic variants in relation to

278 own birthweight.<sup>(46)</sup> However, we do not have genome-wide data in maternal-offspring pairs in UKB  
279 and so cannot explore this here.

## 280 Discussion

281 The possibility that genetic IVs for a specific exposure will associate with many other traits is  
282 increasing as GWAS explore a larger number of SNPs in increasing sample sizes. In this paper we  
283 have described different scenarios that could result in such associations and methods for exploring  
284 where these may cause bias. Beyond confounding by PS, a key concern is attempting to differentiate  
285 vertical from horizontal pleiotropy and using methods to explore and reduce bias from horizontal  
286 pleiotropy. We provide a set of recommendations and demonstrate their use in an applied example  
287 exploring the effect of maternal insomnia on birthweight.

288 This paper brings attention to the pros and cons of hypothesis-driven versus comprehensive  
289 approaches to exploring IV validity. Our motivating example used researchers' knowledge to decide  
290 which non-exposure traits to explore genetic IVs associations with. Specifically, we chose six  
291 observed traits in UKB that we considered plausible causes of offspring birthweight, and as our  
292 analyses in this example shows they reflect plausible horizontal or vertical pleiotropic paths.  
293 However, we have to rely on sensitivity analyses (see Table 2) to control for horizontal pleiotropy via  
294 unexplored traits. This approach is efficient but cannot identify the nature of any unexplored  
295 violation of instrument validity. Sensitivity analyses will identify whether results are likely to be  
296 biased by unbalanced horizontal pleiotropy but if one wanted to explore specific known horizontal  
297 or vertical (mediating) pleiotropy this approach would potentially miss some key paths. A further  
298 limitation is that researchers' knowledge is likely to vary between different researcher groups. An  
299 alternative to *a priori* selecting a defined set of potential pleiotropic traits, is to use a Phenome-wide  
300 search to systematically explore any possible non-exposure traits that associate with our genetic IVs.  
301 This approach has the advantage that it is not limited by researchers' own knowledge and the  
302 variation in this between research groups. Although automated systems for rapid phenome-wide  
303 associations now make a more extensive and systematic approach possible,<sup>(47-49)</sup> there are  
304 challenges in applying our recommendations 3-5 to a possible large number of traits.



305 Of particular importance, when multiple different potential traits (exposure of interest and non-  
306 exposure traits) and the relationship between them is being considered differing measurement error  
307 in each trait may affect the results obtained. In MR and MVMR differing measurement error in  
308 different traits gives the same effect as differing power in each trait and will lead to the effects of  
309 traits measured with more error being less precisely estimated than the effects of those measured  
310 with less error. However, Steiger filtering assumes that each trait is measured with the same error  
311 and can give misleading results for the causal direction between two traits when the true causal  
312 exposure is measured with more error than the true outcome. In our example self-report of  
313 insomnia is likely to be measured with more error than several of the non-exposure traits  
314 considered, in particular maternal height, BMI, age at first birth and education. For these traits  
315 Steiger filtering may misleadingly suggest that the direction of effect is from these traits to insomnia  
316 due to the imprecision in the measurement of insomnia. These issues are relevant to both the use of  
317 prior knowledge to select specific traits to explore as possible pleiotropic paths and to a more  
318 comprehensive and systematic phenome-wide scan approach. However, with the latter there are  
319 many more non-exposure traits where these problems are likely to arise. In our illustrative example  
320 the phenome-wide scan approach identified 478 non-exposure traits associated with one or more of  
321 the 80 insomnia SNPs used in our genetic IV (i.e. 80-fold the six explored on the basis of prior  
322 knowledge). On-the-one hand this suggests we might have missed some key specific pleiotropic  
323 paths, on the other, even with the large sample size used in our example the potential for  
324 uninterpretable imprecise results and possible misleading results is increased with the much larger  
325 number from the phenome-wide scan.

326 Finally, the automated phenome-wide approach is dependent on the nature and quality of the  
327 studies included in the searching tools (e.g. PhenoScanner<sup>(15)</sup> and GWAS Catalog<sup>(16)</sup>) and whilst they  
328 are likely to identify more specific pleiotropic paths than knowledge based approaches, they may  
329 still miss some important paths. Whether researchers decide to focus solely on a limited set of traits  
330 that are known through prior knowledge to influence outcome and could be on a horizontal

331 pleiotropic path or undertake a phenome-wide approach will depend on the specific research  
332 question, including whether that includes an interest in understanding the nature of horizontal  
333 pleiotropic paths or mediation (vertical pleiotropy). It will also depend on available data. A  
334 combination of both could be undertaken with some a prior decision to select a fixed number of the  
335 non-exposure traits identified by the searching tool.

336 Our study provides guidance for further MR studies where genetic IVs were associated with multiple  
337 traits. It may also be relevant to studies using non-genetic IVs (e.g. healthcare practitioner  
338 preference<sup>(50)</sup> or randomization in a randomized control trials<sup>(51)</sup>). In addition to the approaches  
339 outlined here to the situation of identifying that genetic IVs are related to multiple non-exposure  
340 traits, we would recommend triangulating MR results with other methods that have different key  
341 sources of bias to estimate causal effects.<sup>(52)</sup> If results are consistent across such different methods  
342 that increases confidence in the result, even in the presence of remaining concerns about genetic IV  
343 validity.

344 **Funding**

345 This work was supported by the University of Bristol and UK Medical Research Council [grant  
346 numbers, MM\_UU\_00011/1, MM\_UU\_00011/2, MM\_UU\_00011/3 and MM\_UU\_00011/6], the US  
347 National Institute for Health (R01 DK10324), the European Research Council via Advanced Grant  
348 669545, the British Heart Foundation (AA/18/7/34219) and the NIHR Biomedical Research Centre at  
349 University Hospitals Bristol NHS Foundation Trust and the University of Bristol. Q.Y. is funded by a  
350 China Scholarship Council PhD Scholarship (CSC201808060273). M.C.B. is funded by a UK Medical  
351 Research Council Skills Development Fellowship (MR/P014054/1). D.A.L. is an NIHR Senior  
352 Investigator (NF-0616-10102). The funders had no influence on the study design, data collection and  
353 analysis, decision to publish, or preparation of the manuscript.

354 **Acknowledgements**

355 This research was conducted using the UK Biobank Resource under Application Number 23938. The  
356 UK Biobank received ethical approval from the research ethics committee (REC reference for UK  
357 Biobank 11/NW/0382) and participants provided written informed consent.

358 **Competing interests**

359 D.A.L. reports receiving research support from Medtronic and Roche Diagnostics for research  
360 outside the submitted work. All other authors declare no competing interests.

361 **Table 1. Scenarios when an unexpected IV-non-exposure trait association could be observed**

Scenarios	Population stratification (confounding)	Vertical pleiotropy (mediation)	Horizontal pleiotropy	Reverse causality
<b>Directed acyclic graphs<sup>1</sup></b>	<p>1.1  1.2  1.3 </p>	<p>2.1  2.2  2.3 </p>	<p>3.1  3.2  3.3 </p>	<p>4.1 </p>
<b>Violation of assumptions</b>	Yes (1.1, 1.2)/No (1.3)	No	Yes	Yes
<b>Methods to explore the likelihood of the scenario</b>				
<i>One-sample MR with individual data</i>	Check $Z \sim PS$ (e.g. PCs, birthplace, home location or study centre), adjust for PS in Z-W and compare the adjusted estimates with crude estimates	Univariable MR for W-Y, bidirectional MR between X and W and Steiger directionality test (two-sample only), tests for heterogeneity between Z, MR-Egger intercept (two-sample only)		Bidirectional MR between X and Y, Steiger directionality test (two-sample only)
<i>Two-sample MR with summary data</i>	Check genome-wide association studies of X and Y			
<b>Methods to produce valid results</b>				
<i>One-sample MR with individual data</i>	Control for PS in two-stage least squares	Two-stage least squares	Multivariable MR, sisVIVE	Not applicable if Y causes X
<i>Two-sample MR with summary data</i>	Rely on the genome-wide association studies of X and Y	Inverse variance weighted	Multivariable MR, MR-Egger, weighted median, weighted mode, MR-PRESSO, MR-TRYX	
<b>Observed traits in real data example<sup>2</sup></b>	Age at first live birth	Height, body mass index	Age at first live birth, education, ever smoking	Not applicable

362 <sup>1</sup>Z: genetic instrumental variables; X: exposure of interest; Y: outcome of interest; PS: variables representing population stratification; W: non-exposure  
363 traits. For simplicity, the directed acyclic graphs use single nodes even when there may be multiple variables.

364 <sup>2</sup>Our MR found little evidence for an effect of frequency of alcohol intake on birthweight, suggesting it would not bias MR estimates regardless of its causal  
365 relationships with insomnia PRS and insomnia.

366 Abbreviation: MR, Mendelian randomization; PCs, principal components.

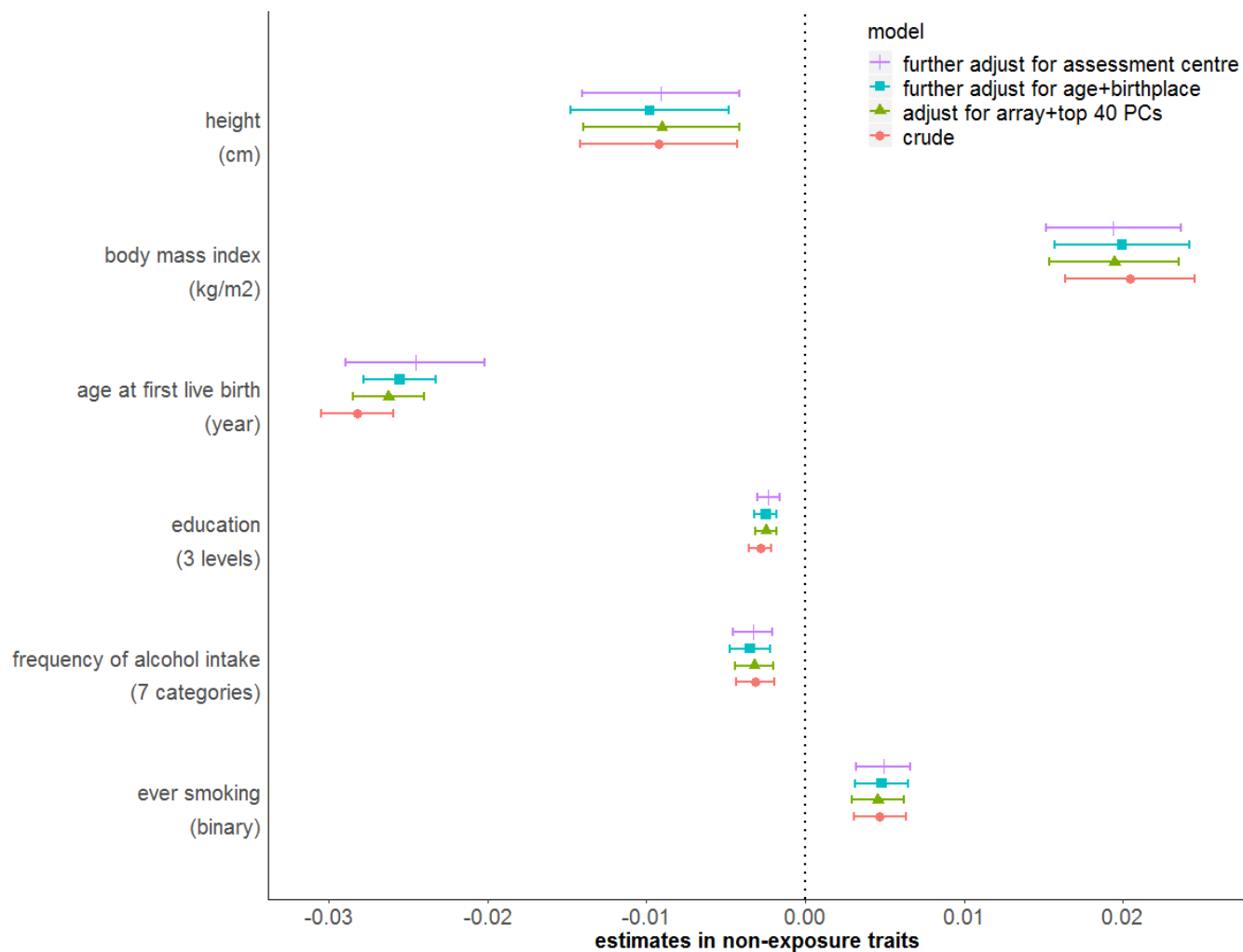
367 **Table 2. Summary of select sensitivity analyses for exploring bias due to horizontal pleiotropy in MR**

Name	Brief description	Assumptions <sup>1</sup>	Other issues
<i>For one-sample MR</i>			
sisVIVE <sup>(29)</sup>	It is an extension to two-stage least squares, which incorporates LASSO penalization.	At least 50% of IVs are valid.	It works for continuous outcomes only, requires a large amount of computer memory, and the current implementation do not provide 95% CIs.
<i>For two-sample MR</i>			
MR-Egger <sup>(30)</sup>	It allows a non-zero intercept to test horizontal pleiotropy.	Instrument strength independent of direct effect	It is sensible to outliers and tends to suffer from low statistical power.
Weighted median <sup>(31)</sup>	It is defined as the median of a weighted empirical density function of the Wald ratio estimates.	At least 50% of weight comes from valid IVs.	Nil
Weighted mode <sup>(32)</sup>	It calculates the weighted mode of the Wald ratio estimates. It will be unbiased even if the majority of SNPs could be invalid but providing the set of SNPs which form the largest homogeneous cluster are valid. <sup>(27)</sup>	Zero modal pleiotropy assumption	Researchers need to choose a bandwidth to obtain the clustering effect, and different bandwidths might provide inconsistent estimates. <sup>(27)</sup>
MR-PRESSO <sup>(17)</sup>	It assesses horizontal pleiotropy based on the contribution of each SNP to heterogeneity and provides adjusted MR estimates by removing outlier SNPs.	Instrument strength independent of direct effect; Outliers (identified via MR-PRESSO global test) are risen due to potential horizontal pleiotropy.	After removing outlier SNPs, the standard errors would decrease. Therefore, it would be more likely to reject the null.
MR-TRYX <sup>(33)</sup>	It assesses horizontal pleiotropy based on the contribution of each SNP to heterogeneity and attempts to adjust for their horizontal pleiotropic effects using extra publicly available GWAS from MR-Base.	Outliers (identified via RadialMR <sup>(53)</sup> ) are risen due to potential horizontal pleiotropy.	GWAS from MR-Base may not cover the whole genome or conducted in the target population (e.g. only female participants).

368 <sup>1</sup>Extra assumptions except for the three key MR assumptions.

369 Abbreviations: CI, confidence interval; GWAS, genome-wide association studies; IV, instrumental variable; MR, Mendelian randomization; SNP, single  
 370 nucleotide polymorphisms.

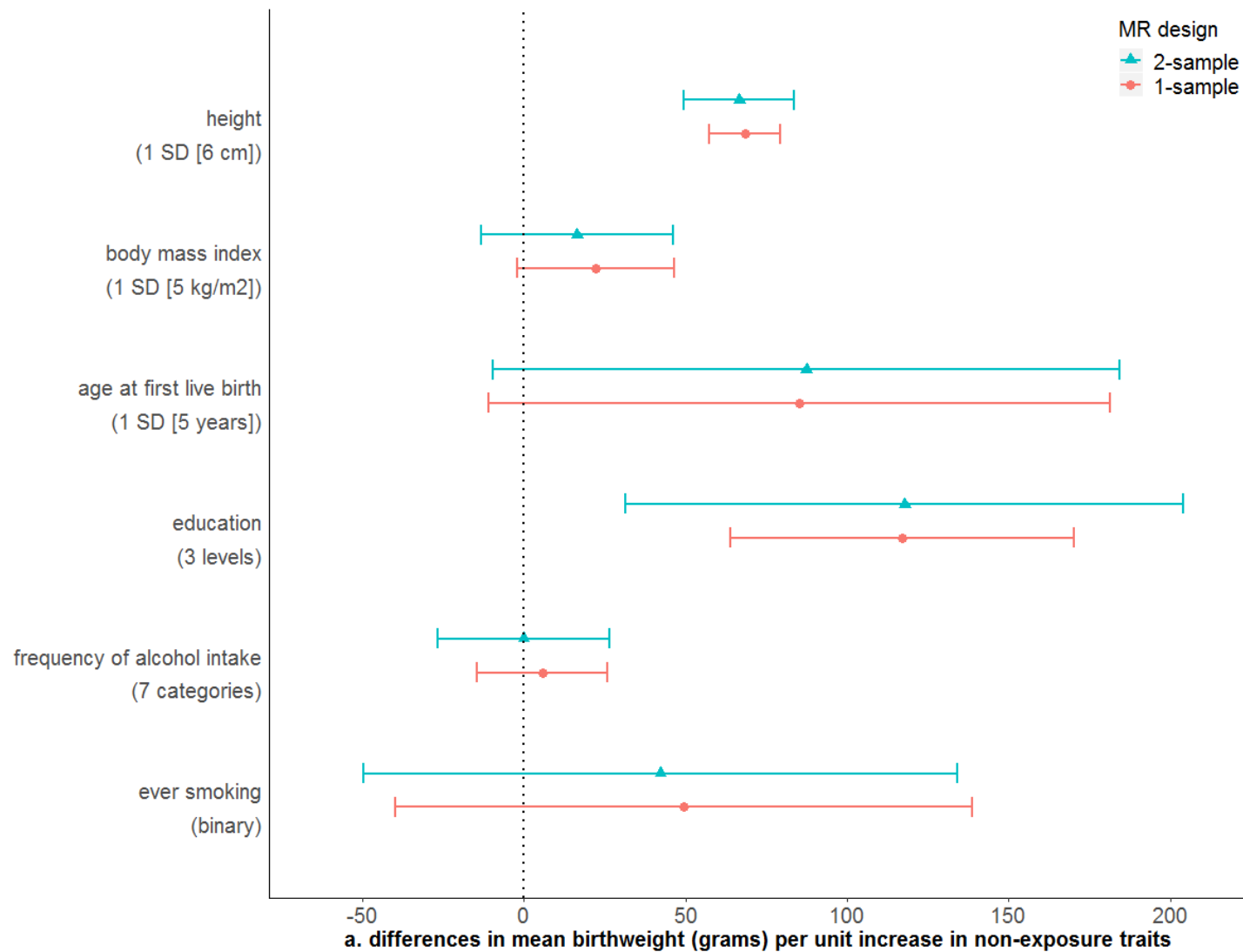
371 **Figure 1. Associations of polygenetic risk score (PRS) for insomnia with six non-exposure traits before and after adjustment for population stratification**



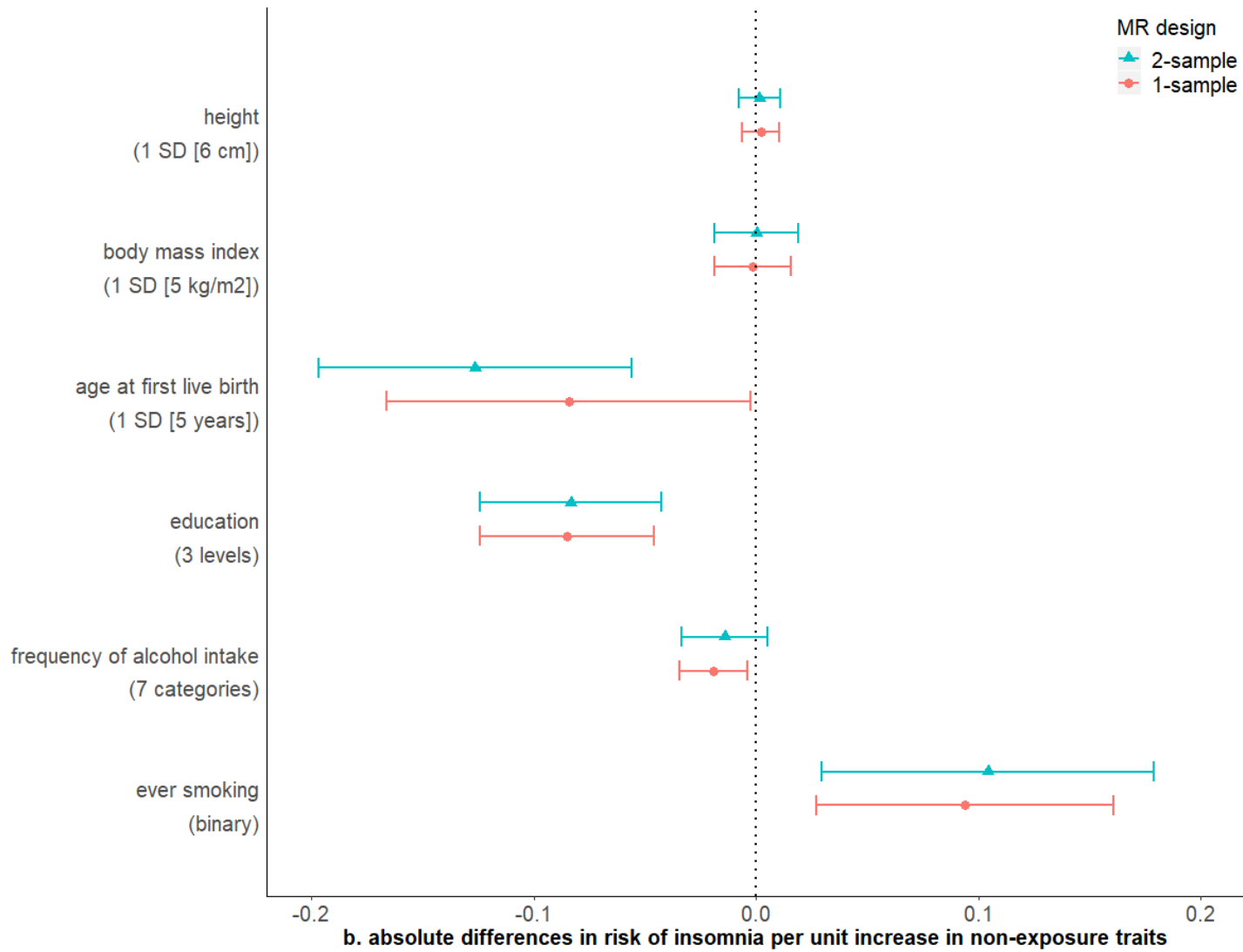
372

373 Estimates are differences in mean non-exposure traits or log odds ratio (ever smoking) per allele increase in PRS. Supplementary Table 1 summarizes how  
 374 education, frequency of alcohol intake and ever smoking are coded in this study.

375 **Figure 2. Mendelian randomization estimates for (a) non-exposure traits-birthweight (W-Y) effects, (b) non-exposure traits-insomnia (W-X) effects, and**  
 376 **(c) insomnia-non-exposure traits (X-W) effects**



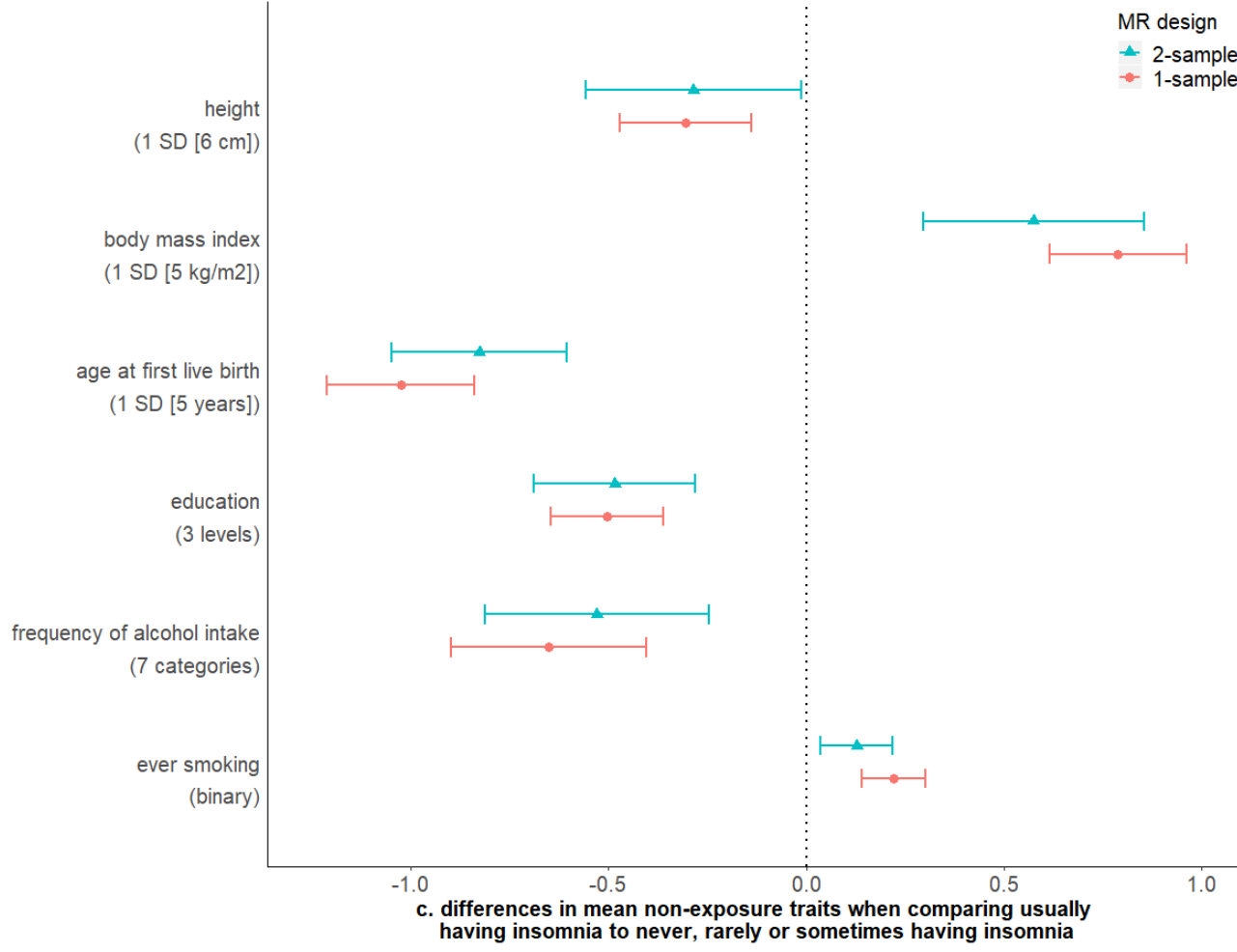
377



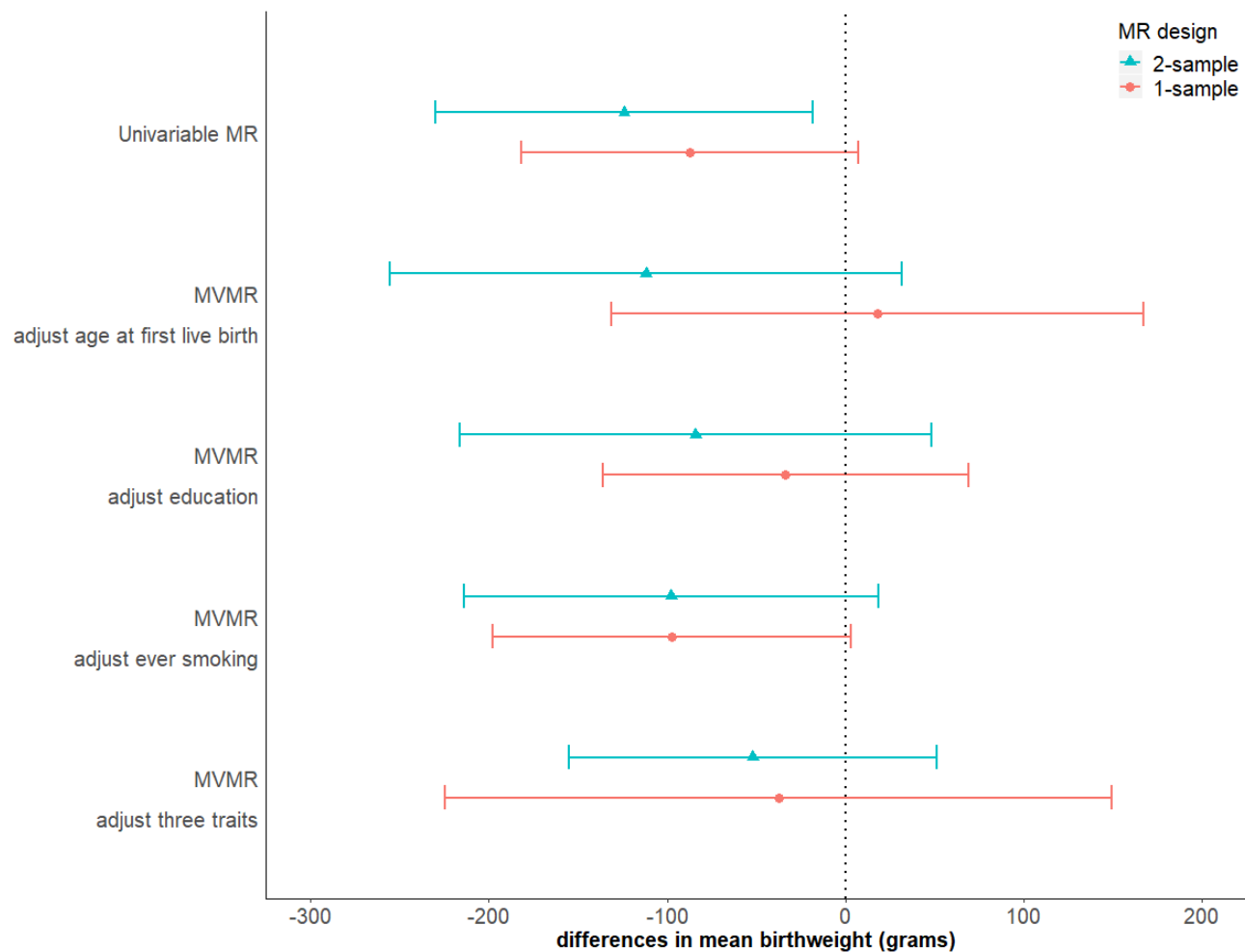
378

379 “Usually” having insomnia is coded as 1, while “sometimes/rarely/never” having insomnia is coded as 0 (Supplementary Table 1).





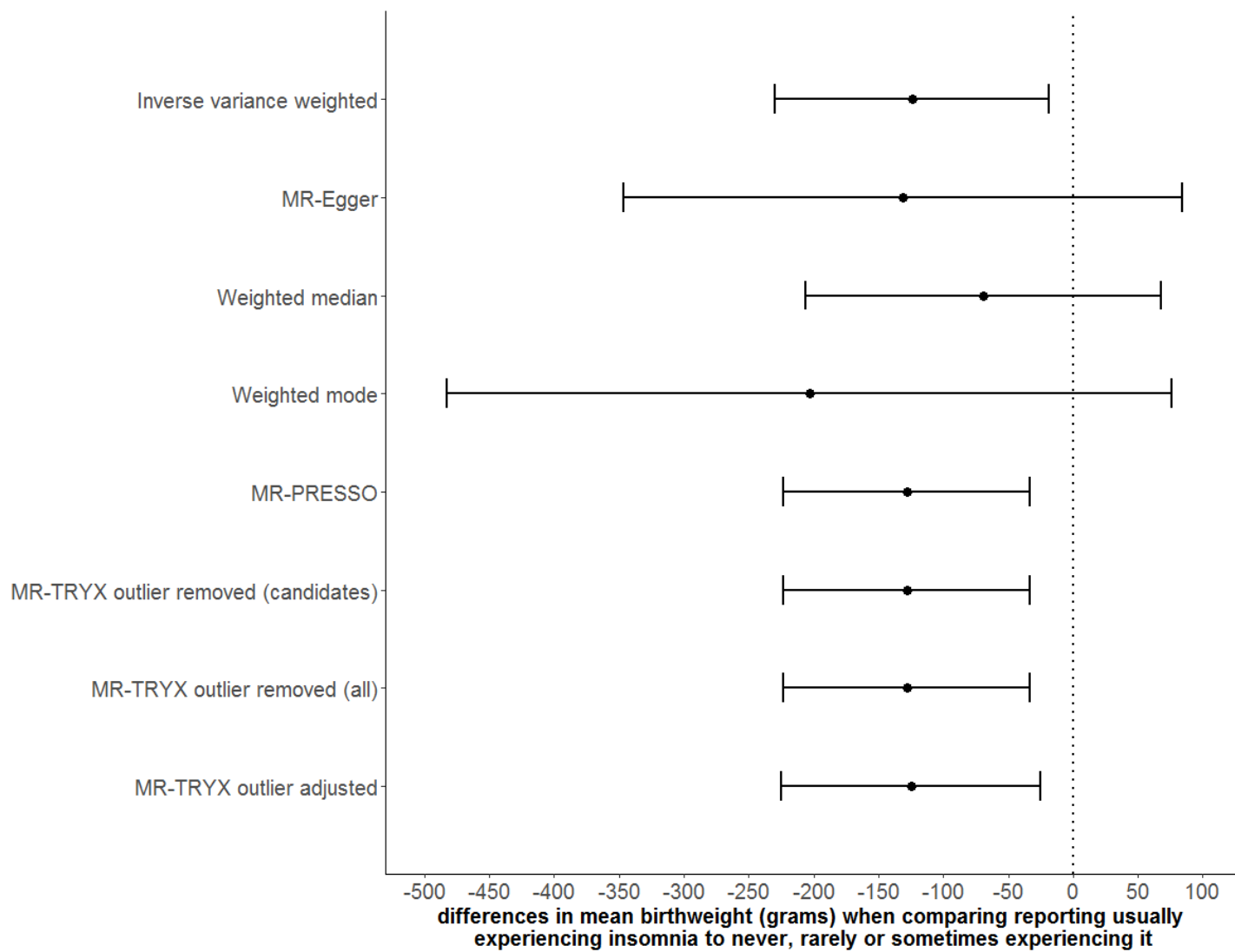
381 **Figure 3. Multivariable Mendelian randomization (MVMR) estimates for the effect of maternal insomnia on offspring birthweight**



382

383 Estimates are differences in mean birthweight when comparing reporting usually experiencing insomnia to never, rarely or sometimes experiencing it with  
384 and without adjustment for potential horizontal pleiotropy via maternal age at first birth, education and ever smoking.

385 **Figure 4. Sensitivity analyses for the effect of maternal insomnia on offspring birthweight using two-sample Mendelian randomization (MR)**



386

## 387 References

- 388 1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute  
389 to understanding environmental determinants of disease? *Int J Epidemiol.* 2003;**32**:1-22.
- 390 2. Lawlor DA, Harbord RM, Sterne JA, *et al.* Mendelian randomization: using genes as  
391 instruments for making causal inferences in epidemiology. *Stat Med.* 2008;**27**:1133-1163.
- 392 3. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide,  
393 glossary, and checklist for clinicians. *Bmj.* 2018;**362**:k601.
- 394 4. Taylor M, Rode L, Bjorngaard J, *et al.* Is smoking heaviness causally associated with alcohol  
395 use? A Mendelian randomization study in four European cohorts. *Int J Epidemiol.* 2018;**47**:1098-  
396 1105.
- 397 5. Budu-Aggrey A, Brumpton B, Tyrrell J, *et al.* Evidence of a causal relationship between body  
398 mass index and psoriasis: A mendelian randomization study. *PLoS Med.* 2019;**16**:e1002739.
- 399 6. Borges MC, Lawlor DA, de Oliveira C, *et al.* Role of Adiponectin in Coronary Heart Disease  
400 Risk: A Mendelian Randomization Study. *Circ Res.* 2016;**119**:491-499.
- 401 7. Richmond RC, Anderson EL, Dashti HS, *et al.* Investigating causal relations between sleep  
402 traits and risk of breast cancer in women: mendelian randomisation study. *BMJ.*  
403 2019;**365**:<https://www.bmj.com/content/365/bmj.l2327>.
- 404 8. Ormel J, Hartman CA, Snieder H. The genetics of depression: successful genome-wide  
405 association studies introduce new challenges. *Transl Psychiatry.* 2019;**9**:114.
- 406 9. Pickrell JK, Berisa T, Liu JZ, *et al.* Detection and interpretation of shared genetic influences on  
407 42 human traits. *Nat Genet.* 2016;**48**:709-717.
- 408 10. Koellinger PD, de Vlaming R. Mendelian randomization: the challenge of unobserved  
409 environmental confounds. *Int J Epidemiol.* 2019.
- 410 11. Haworth S, Mitchell R, Corbin L, *et al.* Apparent latent structure within the UK Biobank  
411 sample has implications for epidemiological analysis. *Nat Commun.* 2019;**10**:333.
- 412 12. Lawlor DA, Wade K, Borges MC, *et al.* A Mendelian randomization dictionary: Useful  
413 definitions and descriptions for undertaking, understanding and interpreting Mendelian  
414 randomization studies [Internet]. *OSF Preprints.* 2019:<https://osf.io/6yzs7/>.
- 415 13. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in  
416 epidemiological studies. *Hum Mol Genet.* 2014;**23**:R89-98.
- 417 14. Richmond RC, Davey Smith G. Commentary: Orienting causal relationships between two  
418 phenotypes using bidirectional Mendelian randomization. *Int J Epidemiol.* 2019.
- 419 15. Staley JR, Blackshaw J, Kamat MA, *et al.* PhenoScanner: a database of human genotype-  
420 phenotype associations. *Bioinformatics.* 2016;**32**:3207-3209.
- 421 16. Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published  
422 genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*  
423 2019;**47**:D1005-d1012.
- 424 17. Verbanck M, Chen CY, Neale B, *et al.* Detection of widespread horizontal pleiotropy in causal  
425 relationships inferred from Mendelian randomization between complex traits and diseases. *Nat*  
426 *Genet.* 2018;**50**:693-698.
- 427 18. Zhu Z, Zheng Z, Zhang F, *et al.* Causal associations between risk factors and common diseases  
428 inferred from GWAS summary data. *Nat Commun.* 2018;**9**:224.
- 429 19. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases  
430 and complex traits. *Nat Genet.* 2018;**50**:1728-1734.
- 431 20. DiPrete TA, Burik CAP, Koellinger PD. Genetic instrumental variable regression: Explaining  
432 socioeconomic and health outcomes in nonexperimental data. *Proc Natl Acad Sci U S A.*  
433 2018;**115**:E4970-e4979.
- 434 21. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and  
435 challenges. *Int J Epidemiol.* 2016;**45**:908-915.
- 436 22. Henry A, Katsoulis M, Masi S, *et al.* The relationship between sleep duration, cognition and  
437 dementia: a Mendelian randomization study. *Int J Epidemiol.* 2019.

- 438 23. Xu S, Gilliland FD, Conti DV. Elucidation of causal direction between asthma and obesity: a  
439 bi-directional Mendelian randomization study. *Int J Epidemiol*. 2019.
- 440 24. Anderson EL, Howe LD, Wade K, *et al*. Education, intelligence and Alzheimer's disease:  
441 Evidence from a multivariable two-sample Mendelian randomization study. *bioRxiv*.  
442 2018:<https://www.biorxiv.org/content/10.1101/401042v401042>.
- 443 25. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely  
444 measured traits using GWAS summary data. *PLoS Genet*. 2017;**13**:e1007081.
- 445 26. Sanderson E, Davey Smith G, Windmeijer F, *et al*. An examination of multivariable Mendelian  
446 randomization in the single-sample and two-sample summary data settings. *Int J Epidemiol*. 2018.
- 447 27. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in  
448 Mendelian randomization studies. *Hum Mol Genet*. 2018;**27**:R195-r208.
- 449 28. Burgess S, Scott RA, Timpson NJ, *et al*. Using published data in Mendelian randomization: a  
450 blueprint for efficient identification of causal risk factors. *Eur J Epidemiol*. 2015;**30**:543-552.
- 451 29. Kang H, Zhang A, Cai TT, *et al*. Instrumental Variables Estimation With Some Invalid  
452 Instruments and its Application to Mendelian Randomization. *Journal of the American Statistical*  
453 *Association*. 2016;**111**:132-144. <https://doi.org/10.1080/01621459.2014.994705>.
- 454 30. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments:  
455 effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;**44**:512-525.
- 456 31. Bowden J, Davey Smith G, Haycock PC, *et al*. Consistent Estimation in Mendelian  
457 Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol*.  
458 2016;**40**:304-314.
- 459 32. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian  
460 randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. 2017;**46**:1985-1998.
- 461 33. Cho Y, Haycock PC, Gaunt TR, *et al*. MR-TRYX: Exploiting horizontal pleiotropy to infer novel  
462 causal pathways. *bioRxiv*. 2018:<https://doi.org/10.1101/476085>.
- 463 34. Morokuma S, Shimokawa M, Kato K, *et al*. Maternal sleep and small for gestational age  
464 infants in the Japan Environment and Children's Study: a cohort study. *BMC Res Notes*. 2017;**10**:394.
- 465 35. Warland J, Dorrian J, Morrison JL, *et al*. Maternal sleep during pregnancy and poor fetal  
466 outcomes: A scoping review of the literature with meta-analysis. *Sleep Med Rev*. 2018;**41**:197-219.
- 467 36. Collins R. What makes UK Biobank special? *Lancet*. 2012;**379**:1173-1174.
- 468 37. Jansen PR, Watanabe K, Stringer S, *et al*. Genome-wide analysis of insomnia in 1,331,010  
469 individuals identifies new risk loci and functional pathways. *Nat Genet*. 2019;**51**:394-403.
- 470 38. Randall JC, Winkler TW, Kutalik Z, *et al*. Sex-stratified genome-wide association studies  
471 including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS*  
472 *Genet*. 2013;**9**:e1003500.
- 473 39. Locke AE, Kahali B, Berndt SI, *et al*. Genetic studies of body mass index yield new insights for  
474 obesity biology. *Nature*. 2015;**518**:197-206.
- 475 40. Barban N, Jansen R, de Vlaming R, *et al*. Genome-wide analysis identifies 12 loci influencing  
476 human reproductive behavior. *Nat Genet*. 2016;**48**:1462-1472.
- 477 41. Okbay A, Beauchamp JP, Fontana MA, *et al*. Genome-wide association study identifies 74 loci  
478 associated with educational attainment. *Nature*. 2016;**533**:539-542.
- 479 42. Nealelab. We're thrilled to announce an updated GWAS analysis of the UK Biobank.  
480 2018:<http://www.nealelab.is/uk-biobank>.
- 481 43. Millard LAC, Davies NM, Tilling K, *et al*. Searching for the causal effects of body mass index in  
482 over 300 000 participants in UK Biobank, using Mendelian randomization. *PLoS Genet*.  
483 2019;**15**:e1007951.
- 484 44. Zhang G, Bacelis J, Lengyel C, *et al*. Assessing the Causal Relationship of Maternal Height on  
485 Birth Size and Gestational Age at Birth: A Mendelian Randomization Analysis. *PLoS Med*.  
486 2015;**12**:e1001865.
- 487 45. Tyrrell J, Richmond RC, Palmer TM, *et al*. Genetic Evidence for Causal Relationships Between  
488 Maternal Obesity-Related Traits and Birth Weight. *Jama*. 2016;**315**:1129-1140.

- 489 46. Warrington NM, Beaumont RN, Horikoshi M, *et al.* Maternal and fetal genetic effects on  
490 birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet.* 2019;**51**:804-814.
- 491 47. Hemani G, Zheng J, Elsworth B, *et al.* The MR-Base platform supports systematic causal  
492 inference across the human phenome. *Elife.* 2018;**7**.
- 493 48. Millard LA, Davies NM, Timpson NJ, *et al.* MR-PheWAS: hypothesis prioritization among  
494 potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci*  
495 *Rep.* 2015;**5**:16645.
- 496 49. Zheng J, Haberland V, Baird D, *et al.* Phenome-wide Mendelian randomization mapping the  
497 influence of the plasma proteome on complex diseases. *bioRxiv.* 2019.
- 498 50. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the  
499 estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat.*  
500 2007;**3**:Article 14.
- 501 51. Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for  
502 treatment contamination in randomised controlled trials. *Bmj.* 2010;**340**:c2073.
- 503 52. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J*  
504 *Epidemiol.* 2016;**45**:1866-1886.
- 505 53. Bowden J, Spiller W, Del Greco MF, *et al.* Improving the visualization, interpretation and  
506 analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial  
507 regression. *Int J Epidemiol.* 2018;**47**:2100.
- 508