

1 **Genome-wide analyses of behavioural traits biased by misreports and longitudinal changes**

2

3 Angli Xue¹, Longda Jiang¹, Zhihong Zhu¹, Naomi R. Wray^{1,2}, Peter M. Visscher¹, Jian Zeng¹, Jian
4 Yang^{1,3,*}

5

6 ¹ Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072,
7 Australia

8 ² Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, 4072, Australia

9 ³ Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

10 * Correspondence: Jian Yang (jian.yang.qt@gmail.com)

11

12 **Abstract**

13 Genome-wide association studies (GWAS) have discovered numerous genetic variants associated
14 with human behavioural traits. However, behavioural traits are subject to misreports and longitudinal
15 changes (MLC) which can cause biases in GWAS and follow-up analyses. Here, we demonstrate that
16 individuals with higher disease burden in the UK Biobank ($n = 455,607$) are more likely to
17 misreport or reduce their alcohol consumption (AC) levels, and propose a correction procedure to
18 mitigate the MLC-induced biases. The AC GWAS signals removed by the MLC corrections are
19 enriched in metabolic/cardiovascular traits. Almost all the previously reported negative estimates of
20 genetic correlations between AC and common diseases become positive/non-significant after the
21 MLC corrections. We also observe MLC biases for smoking and physical activities in the UK
22 Biobank. Our findings provide a plausible explanation of the controversy about the effects of AC on
23 health outcomes and a caution for future analyses of self-reported behavioural traits in biobank data.

24 **Introduction**

25 Behaviours and lifestyles are modifiable risk/protective factors for common diseases in humans. In
26 the past few decades, one of the most controversial debates in public health is on the effect of alcohol
27 consumption (AC) on common diseases, especially cardiovascular/metabolic diseases. Large-scale
28 meta-analyses of epidemiological studies^{1,2} on AC concluded that “no level of alcohol consumption
29 improves health”³. This conclusion, however, is contradictory to the negative estimates of genetic
30 correlation (r_g) between AC and several diseases such as obesity⁴⁻⁷, major depressive disorder
31 (MDD)^{6,7}, Parkinson’s disease⁵ and type 2 diabetes (T2D)⁵ reported in recent genome-wide
32 association studies (GWAS) and also contradictory to the protective effects of moderate drinking
33 reported in observational studies^{8,9}. Different hypotheses have been proposed to explain these
34 discrepancies, including: 1) heavy AC might alter metabolism or impair nutrient absorption^{10,11},
35 meaning that the effect is dosage-dependent; 2) people who had health problems may quit or reduce
36 drinking, or underreport their intake level¹²; and 3) some other common explanations include
37 confounding factors⁶ (e.g., socio-economic status and physical activities) and collider bias^{13,14}.
38 Nevertheless, to date, no study has provided in-depth investigation into the causes of the
39 discrepancies.

40
41 In epidemiological or genetic studies, phenotypic data of behavioural and lifestyle traits are often
42 collected from self-reported questionnaires, which are subject to misreporting (i.e., self-report biases),
43 especially for questions related to smoking, drinking and drug use¹⁵⁻¹⁸. These behaviours are also
44 subject to change during lifetime¹⁹⁻²², for instance in response to disease diagnosis, but data to track
45 such longitudinal variations are rarely available. Both misreports and longitudinal changes (hereafter
46 referred to as MLC) could change the distribution of behavioural phenotypes and thus may affect the
47 results of both epidemiological and genetic studies.

48
49 In this study, we set out to investigate biases due to MLC in genetic analyses of self-reported
50 behavioural traits including AC, tobacco smoking, and physical activities in the UK Biobank
51 (UKB)²³. The UKB includes detailed questionnaires of these behavioural traits, providing a unique
52 resource to investigate the potential pitfalls in the analyses of self-reported phenotypes. We
53 demonstrate that MLC could induce biases in GWAS for these traits and follow-up analyses that use
54 summary statistics from the GWAS. We then propose a correction procedure to mitigate the MLC
55 biases and elaborate on why some of previous studies might suffer from MLC.

56 57 **Results**

58 **Misreports and longitudinal changes in alcohol consumption**

59 Misreports are common in self-reported data sets¹⁵⁻¹⁸ but often overlooked in genetic analyses. Here,
60 we focused on the analyses of AC because 1) its relationship with common diseases is controversial;

61 2) the data required by our investigations and corrections are available; and 3) the sample size is large
62 ($n = 455,607$). In this study, our definition of misreports for AC includes misreporting about
63 drinking status²⁴, underreporting the AC level^{15,17}, and selective recall of the question about AC
64 level²⁵, all of which might occur in the UKB. These kinds of misreports are mainly attributed to^{26,27}
65 social desirability^{28,29} (i.e., the tendency of participants to answer questions in ways that make them
66 viewed favourably by others) and recall bias^{30,31} (i.e., the accuracy and completeness of past events
67 recalled by participants are influenced by subsequent events that they experienced). First, 14,488
68 UKB participants identified themselves as never drinkers, but combining additional data from follow-
69 up questionnaires and medical records³² revealed that at least 10% of these individuals were very
70 likely to have drinking history, e.g., previously diagnosed as having alcoholic hepatitis or alcohol use
71 disorder (**Supplementary Note 1**). This means that classifying self-reported never drinkers as lifetime
72 abstainers could be problematic²⁴. Thus, our analyses on AC were mainly focused on current drinkers
73 unless specified elsewhere. Second, 9,064 individuals are classified as current drinkers but reported
74 zero consumption level, indicating possible underreporting. Third, 66,058 individuals (15.6% current
75 drinkers) reported their alcohol intake frequency and other related questions but did not report their
76 AC levels, suggesting a potential selective recall bias. It has been shown previously that heavy
77 drinkers tend to be unresponsive²⁵, and a high non-response rate could lead to an underestimate of the
78 average AC level in the sample³³. To investigate the characteristics of the suspected misreporting
79 individuals, we examined the phenotypes of 18 common diseases in the UKB and used disease count
80 (the number of diseases carried) as an indicator of disease burden for each participant (**Methods**,
81 **Table 1**, and **Supplementary Table 1**). We observed that unresponsive individuals had a much
82 higher mean disease count than individuals with complete responses (1.63 vs. 1.37, Welch t-test $P =$
83 6.35×10^{-294} ; **Table 1**). The suspected underreporting individuals ($n = 9,064$) also showed a higher
84 mean disease count than the remaining current drinkers (1.73 vs. 1.36, Welch t-test $P =$
85 2.68×10^{-87}).

86

87 Another important source of bias is the change in drinking behaviour during the life course for
88 reasons such as changes in health status. For instance, if people change their AC level because they
89 are affected by a disease, such a disease ascertainment will give rise to a bias in observed or genetic
90 relationship between AC and the disease. In the UKB, all the current drinkers ($n = 424,507$) were
91 asked a question “compared to 10 years ago, do you drink less/the same/more nowadays?”
92 (**Methods**), and ~62% of them reported “less” or “more”. We denote the three groups of individuals
93 as LESS, SAME and MORE, respectively. The LESS group ($n = 191,653$) had a lower average AC
94 level, higher disease prevalence for several common diseases, and higher mean disease count than
95 individuals in the other two groups (**Table 1** and **Supplementary Table 1**). A follow-up question was
96 to ask the participants to choose the reason(s) why they reduced drinking, and the available options

97 include illness, health precaution, and financial reasons (**Table 2**). There were 15,889 individuals
98 choosing illness/ill health or doctor's advice as the primary reason for reducing drinking, and their
99 mean disease count was nearly twice that of all other current drinkers (**Table 2**). In the subgroup of
100 individuals who reported AC and had reduced drinking due to illness or doctor's advice ($n = 11,886$),
101 the prevalence of cardiovascular disease (CVD) was 0.411, ~2.7 times higher than that in all current
102 drinkers (0.154), providing strong evidence of disease ascertainment of AC (**Supplementary Table**
103 **1**).

104

105 **Biases in GWAS for alcohol consumption due to MLC**

106 We conducted GWAS analyses for AC with and without correcting for MLC. The MLC corrections
107 included excluding individuals who might underreport AC level, excluding individuals who reduced
108 drinking due to illness/doctor's advice, and adjusting the mean and variance difference in three
109 longitudinal change groups (**Methods** and **Supplementary Figures 1-2**). There were 53 and 47
110 independently genome-wide significant loci ($P_{\text{GWAS}} < 5 \times 10^{-8}$) before and after the corrections,
111 respectively (**Supplementary Table 2** and **Supplementary Figure 3**). We identified 16 loci that
112 became non-significant after the corrections ($P_{\text{GWAS}} \geq 5 \times 10^{-8}$, **Supplementary Table 2**). By
113 searching the top associated SNPs at these loci in an online database PheWAS³⁴ (**URLs**), we found
114 that 44.9% of associated phenotypes ($P_{\text{PheWAS}} < 5 \times 10^{-8}$) were metabolic/cardiovascular traits
115 such as body mass index (BMI), triglyceride (TC), and coronary artery disease (CAD) (**Figure 1**). We
116 showed by a down-sampling analysis that the number of loci that became non-significant after the
117 MLC corrections was significantly larger than that expected from a loss of sample size, and that 10
118 loci that became genome-wide significant after the MLC corrections were likely to be masked by
119 MLC in the uncorrected GWAS (**Methods, Supplementary Table 2-3**). These results were in line
120 with the simulation results (**Methods** and **Supplementary Note 2**) that MLC could reduce the power
121 to detect true signals and induce spurious signals due to disease ascertainment (**Supplementary**
122 **Figures 3-9**).

123

124 **Estimates of genetic correlation biased by MLC**

125 Biases in GWAS data due to MLC are expected to carry over to follow-up analyses using summary
126 statistics of the GWAS, such as the genetic correlation (r_g) analysis. To demonstrate such biases, we
127 estimated r_g between AC and 18 common diseases in the UKB by the bivariate LD score regression³⁵
128 (LDSC) (**Methods**) using AC GWAS data from each of the three longitudinal change groups or the
129 whole sample. Before the MLC corrections, we observed substantial differences between \hat{r}_g
130 (between AC and diseases) estimated using AC GWAS data from the LESS, SAME and MORE
131 groups (**Table 1**). We also estimated r_g between different AC GWAS data sets (**Supplementary**
132 **Tables 5-6**, and **Supplementary Figure 10**) and found that the \hat{r}_g between AC in the LESS and

133 MORE groups was significantly different from 1 ($\hat{r}_g = 0.796$, standard error (s.e.) = 0.074). All
134 these results suggest that there was heterogeneity between AC data from the three longitudinal change
135 groups. The heterogeneity was also demonstrated in an additional analysis where we estimated r_g
136 between AC (using data from the UKB) and 234 traits (using data from LD-Hub³⁶) and found that the
137 \hat{r}_g using AC GWAS data from the LESS group were substantially different from those using AC
138 GWAS data from the MORE group, with more than half of the \hat{r}_g (143/234) in the opposite direction
139 between the two groups (**Supplementary Figure 11**). Notably, after the MLC corrections, the \hat{r}_g
140 between all pairwise AC GWAS data sets were close to 1 (ranging from 0.91 to 0.99; **Supplementary**
141 **Table 6**).

142

143 In the r_g analysis using AC GWAS data from the whole sample without the MLC corrections, AC
144 showed nominally significant ($p < 0.05$) negative \hat{r}_g with 3 diseases (i.e., T2D, hypertensive disease
145 and iron deficiency anemias) (**Figure 2** and **Supplementary Table 4**). Negative estimates of r_g
146 between AC and diseases have also been reported in the literature. For instance, Clarke et al.⁴ show a
147 negative \hat{r}_g between AC and obesity, and Liu et al.⁵ show that AC is negatively genetically correlated
148 with several common diseases including Parkinson's disease, obesity, and T2D. However, after the
149 MLC corrections, AC showed nominally significant \hat{r}_g with 8 diseases, all of which were positive
150 (**Figure 2**). These results imply that the negative estimates of r_g between AC and diseases from the
151 analyses without the MLC corrections (including those in prior works) were caused by disease
152 ascertainment. Nevertheless, this conclusion is not definitive because the ground truth is unknown in
153 real data analysis. Hence, we sought to verify it by simulation (**Methods** and **Supplementary Note**
154 **2**), and the results showed that the estimated SNP effect correlation (\hat{r}_b) between a simulated
155 phenotype and disease gradually changed to the opposite direction as the strength of disease
156 ascertainment increased (**Supplementary Note 2** and **Supplementary Figures 5** and **7**), supporting
157 our conclusion. The simulation results also showed that after MLC corrections, r_b was slightly
158 underestimated but with no bias in direction in the presence of disease ascertainment
159 (**Supplementary Figure 9**).

160

161 Socio-economic status (SES) has been shown to affect people's alcohol use and health outcomes, and
162 several studies have shown that people with higher SES tend to have higher AC levels and lower
163 disease risks than people with lower SES^{37,38}. Clarke et al.⁴ and Liu et al.⁵ show positive \hat{r}_g between
164 AC and educational attainment (EA). We observed a similar estimate in our study ($\hat{r}_g = 0.082$, $s. e. =$
165 0.021) before the MLC corrections, but the estimate became non-significant after the MLC
166 corrections (**Figure 3**), likely because MLC are associated with EA. For example, the mean years of
167 schooling of individuals who reduced AC due to illness/doctor advice (12.76, $s. e. m. = 0.05$) was

168 significantly lower than that of the remaining current drinkers (14.23, *s. e. m.* = 0.01). We also
169 included household income (HI) and social deprivation (SD) in the analysis (**Methods**), and our
170 results showed that $\hat{\tau}_g$ between AC and HI or SD could also be biased by MLC (**Supplementary**
171 **Table 9**).

172

173 **Estimates of causal effect biased by MLC**

174 Mendelian randomization (MR) is a method that uses genetic variants as instrumental variables to
175 infer causal relationship between exposure and outcome^{39,40}. As the MR analysis relies on GWAS
176 data, it may also be affected by the MLC biases as described above. We used BMI in the UKB as an
177 example to demonstrate the performance of MR in the presence of MLC, based on several commonly
178 used MR methods including IVW (inverse variance weighted)⁴¹, MR-Egger⁴², GSMR⁴³, weighted
179 median⁴⁴. While the estimates from weighted median, MR-Egger and GSMR were all positive and
180 consistent across all the analyses with or without the MLC corrections, the estimates from IVW,
181 simple median⁴⁴ and MR-PRESSO⁴⁵ seemed to be sensitive to MLC with some of them being
182 negative (**Figure 4**). The negative estimates from the analyses without the MLC corrections were
183 likely to be driven by the 16 loci that were removed by the MLC corrections (note that the mean per-
184 SNP MR estimate at the 16 loci was -0.077). GSMR provided the most robust estimates among the
185 MR methods tested (**Figure 4** and **Supplementary Table 10**) because it has a step (called HEIDI-
186 outlier filtering) that can identify and remove some of the genetic instruments whose effects were
187 biased by pleiotropy/confounding, diminishing the bias in the estimated causal effects
188 (**Supplementary Figure 14**). Notably, after the MLC corrections of AC, the estimates from all the
189 MR methods were largely consistent (**Supplementary Table 10**). We also ran the GSMR analysis in
190 a reverse direction without the MLC corrections and found significant negative effect of BMI on AC
191 ($\hat{b}_{BMI \rightarrow AC} = -0.076, P = 1.11 \times 10^{-33}$) (**Supplementary Table 11**), consistent with the
192 observation above that high BMI might be one of reasons to reduce AC (**Table 1**).

193

194 In addition to the UKB data, we also analysed GWAS summary data for AC from Liu et al.⁵ with a
195 sample size of ~1 million consisting of ~42.9% of the sample from 23andMe and ~33.0% from the
196 UKB (**Methods**). The results were similar to those from the analyses above using the AC GWAS data
197 from the UKB only (**Supplementary Table 10** and **Supplementary Figure 15**), which implies that
198 MLC may not be UKB-specific but also exist in other data sets because otherwise the biases would be
199 smaller in this analysis given only one-third of the AC data were from the UKB. We further
200 confirmed the biases in MR analyses from MLC by simulation (**Supplementary Figures 6** and **8**),
201 and demonstrated that the estimates of causal effects from MR were nearly unbiased after the MLC
202 corrections (**Supplementary Figure 9**).

203

204 **The J-shaped relationship between AC and CVD**

205 In epidemiological studies, there are debates about whether moderate drinking is protective against
206 CVD because of an observed J-shaped relationship between AC and CVD^{2,19,46}. We showed that the
207 moderate drinking group ($0 < AC \leq 25$ gram/week), often used as the reference group to compute the
208 effect (odds ratio, OR) of AC on disease risk, was enriched with individuals from the LESS group
209 which had a higher CVD incidence than the SAME and MORE groups (**Supplementary Figure 12**).
210 This could result in a higher CVD incidence in the reference group than average, leading to a J-shaped
211 relationship between AC and CVD (**Supplementary Figure 13A**). Although the J-shaped relationship
212 between AC and CVD did not change much after MLC corrections (**Supplementary Figure 13B**), it
213 became monotonically increasing after excluding the LESS group from the reference
214 (**Supplementary Figure 13C**). Polygenic predictor of AC showed no evidence for any protective
215 effect of moderate drinking against CVD (**Supplementary Figure 13D**), consistent with the result
216 from a previous study⁴⁶. Our results indicated that the J-shaped relationship between AC and CVD
217 observed in epidemiological studies might be driven by disease ascertainment (**Supplementary Note**
218 **3**).

219

220 **Biases from MLC in other self-reported behavioural traits**

221 Self-reported smoking data in the UKB is also likely to suffer from MLC. Similar to that for AC, all
222 the current smokers were asked “Compared to 10 years ago do you smoke less/the same/more
223 nowadays?”. We partitioned the current smokers ($n = 32,801$) into the LESS, SAME, and MORE
224 groups (**Supplementary Note 4**). The LESS group had a higher disease count (1.69, *s. e. m.* = 0.01)
225 than the SAME group (1.56, *s. e. m.* = 0.01) but a lower disease count than the MORE group (1.84,
226 *s. e. m.* = 0.03) (**Supplementary Table 12**); these results were different from those observed in AC.
227 In the LESS group, individuals who had reduced CPD because of illness/doctor’s advice had nearly
228 twice the disease count (2.73, *s. e. m.* = 0.03) compared to the mean in the entire sample (1.45,
229 *s. e. m.* = 0.002) or all current smokers (1.66, *s. e. m.* = 0.01), indicating that smoking intensity was
230 also ascertained by disease burden. However, unlike AC, the \hat{r}_g between CPD and common diseases
231 were mostly consistent across the LESS, SAME and MORE groups (**Supplementary Figure 16**), and
232 there were negligible differences between the \hat{r}_g estimated using the CPD GWAS data of the whole
233 sample before and after correcting for MLC (**Methods; Supplementary Table 13** and
234 **Supplementary Figure 16**). These results are partly due to the difference in longitudinal change
235 pattern between AC and CPD in the UKB (**Supplementary Tables 1** and **14**). For example, the mean
236 AC in the subgroup of individuals who reduced AC due to illness (7.33 units/week) was lower than
237 that in the entire LESS group (8.56 units/week) but with a significantly higher mean disease count,
238 resulting in a negative correlation between AC and diseases (**Table 2** and **Supplementary Table 12**),

239 whereas the mean CPD in the subgroup due to illness (12.28 cigarettes/day) was higher than that in
240 the entire LESS group (11.70 cigarettes/day) (**Supplementary Table 12**).

241

242 Finally, we investigated physical activities (PA) in the UKB. The PA traits included self-reported
243 METT (Metabolic Equivalent Task in Total) scores, IPAQ (International Physical Activity
244 Questionnaires - short form, **URLs**), and overall acceleration average (OAA, measured by wrist-worn
245 accelerometers). IPAQ is a derived categorical trait (low, moderate, and high) that utilizes information
246 from the METT and its three subsets: walking, moderate, and vigorous activities (**Methods, URLs**).
247 We first estimated r_g between METT, IPAQ and OAA and between METT from the three IPAQ
248 subgroups (**Supplementary Figure 17**). We found a significant genetic heterogeneity between METT
249 and IPAQ ($\hat{r}_g = 0.795$, $s. e. = 0.016$) and a small genetic overlap of either METT or IPAQ with
250 OAA ($\hat{r}_g = 0.232$ with $s. e. = 0.037$ for METT and $\hat{r}_g = 0.390$ with $s. e. = 0.034$ for IPAQ). We
251 then estimated r_g between PA and 18 common diseases. While the \hat{r}_g of IPAQ and OAA with the
252 diseases were mostly negative, METT showed positive \hat{r}_g with most diseases (**Supplementary Table**
253 **15 and Supplementary Figure 18**). We also found that the \hat{r}_g of METT from the low IPAQ subgroup
254 with the diseases were highly consistent with those of IPAQ and OAA but mostly in the opposite
255 direction to those of METT from the moderate and high IPAQ subgroups (**Supplementary Figure**
256 **18**), indicating potential biases in METT from the moderate and high IPAQ subgroups, in line with
257 the finding from a previous study⁴⁷. In addition, the phenotypic correlation between the first and third
258 assessment ($n = 11,484$) of METT was only 0.431, implying substantial longitudinal changes.
259 Unfortunately, these changes are undocumented for the majority of UKB participants. Given that the
260 \hat{r}_g of IPAQ with diseases were highly consistent (in direction) with those of OAA (**Supplementary**
261 **Figure 18**), IPAQ appears to be a more robust indicator for PA than METT.

262

263 Discussion

264 In this study, we raised concerns that genetic analyses of human behavioural traits could be biased by
265 misreports and longitudinal changes. AC in the UKB was used as the main example to demonstrate
266 the detrimental effects of MLC on several genetic analyses commonly used to identify variant-trait
267 associations or estimate genetic or causative relationship between traits. Our results showed that
268 disease ascertainment was likely to be a main cause of the MLC biases, which can be largely
269 corrected for using additional information (e.g., intake frequency and medical records) and coarse
270 longitudinal data (e.g., self-reported longitudinal change direction). Our results also showed that the
271 MLC corrections procedure proposed in this study added value to the routine quality controls in
272 GWAS for behavioural traits like AC. Additionally, biases due to longitudinal changes appeared to be
273 larger than that due to misreports, because the longitudinal changes were observed in more than half
274 of the participants, while misreports only accounted for 10~20% of the UKB sample (at least for those

275 we have identified thus far), as verified in our simulations (**Supplementary Note 2** and
276 **Supplementary Figures 5-8**).

277

278 Our findings provide a plausible explanation for the long-standing controversy about the effects of
279 AC on health outcomes in genetic^{4,5} and epidemiological studies^{2,3,19}. While it seems that most
280 inconsistent estimates in previous studies were due to MLC, there are several reasons why some
281 studies suffered from stronger biases than others. First, the average AC level varied across datasets
282 (from 2.9 to 19.3 units/week across 24 studies)⁵, suggesting heterogeneity in drinking behaviours
283 among different regions/populations. Second, since we have demonstrated that biases from MLC were
284 mainly attributable to disease ascertainment, difference in disease prevalence between populations
285 may lead to different patterns of MLC. Third, as MLC are related to other factors, such as SES,
286 adjusting those factors for AC may offset some of the biases in one way or another. Last but not least,
287 MLC could vary for different age or sex groups. For instance, disease ascertainment is expected to
288 have a larger influence in middle-aged populations than in younger populations because younger
289 populations are less likely to be affected by common diseases investigated in this study⁴⁸. This is
290 supported by the observation that the older UKB participants had a higher mean disease count with a
291 higher proportion of them reducing AC due to illness/doctor's advice (**Supplementary Figure 19**).
292 We also observed differences in the male/female ratio between the three groups (0.551 in the LESS
293 group, 0.488 in the SAME group, and 0.373 in the MORE group were male). Note that we fitted age
294 and sex as covariates in GWAS so the biases due to age or sex are likely to be limited.

295

296 Our study certainly has limitations as it is almost impossible to correct for all the biases with limited
297 availability of relevant data. First, the 9,064 individuals who were suspected to underreport their AC
298 are very likely to be only a subset of all the underreporting individuals. Thus, more effective methods
299 are needed to identify the remaining underreporting individuals. Second, ~58% of the individuals with
300 reduced alcohol intake reported that the reduction was due to "other reasons" or "do not know" in the
301 survey (**Table 2**). If some of these individuals reduced AC also because of illness, then not taking that
302 into account could retain a bias in the analysis. Third, some participants may have misreported their
303 longitudinal changes, leading to an incorrect classification of longitudinal change groups. Fourth, the
304 coarse longitudinal change information itself is cross-sectional (10 years before the time point of the
305 first assessment), meaning that some of the changes occurred beyond the time frame might not be
306 accounted for in this study. Last but not the least, 15% of the current drinkers who did not report their
307 AC level were removed from the analysis. One solution, as implemented in a previous study⁴⁹, is to
308 impute the missing values based on intake frequency and gender. However, 99.8% of the
309 nonresponsive individuals in the UKB are occasional drinkers while this proportion is only 9.4% for
310 the responsive individuals, which might lead to a systematic error between the observed and imputed
311 data sets.

312

313 In conclusion, we advise awareness of the pitfalls when analysing data on behavioural traits in
314 biobank data sets such as the UKB. Misreports and longitudinal changes of behavioural traits by
315 disease ascertainment could create biases and thereby induce spurious signals and a loss of power in
316 GWAS. Biases in GWAS summary statistics due to MLC could further lead to biased estimates in
317 follow-up analyses such as genetic correlation and Mendelian randomization. As more biobank data
318 sets have become accessible, it is important to identify, investigate and correct for these biases in all
319 kinds of behavioural traits including smoking, drinking, diet, physical activity, sleep, and self-rated
320 status. A longitudinal study of 1 million individuals for several decades seems impractical at present,
321 but we have shown that the biases in AC can be largely corrected for by phenotypic QC and
322 longitudinal adjustment when additional phenotype information (intake frequency, medical records,
323 longitudinal change and reasons, etc.) are available. Questionnaires on lifetime use may provide more
324 accurate estimates of the effects of behaviours on health outcomes at a much lower cost than a
325 longitudinal follow-up study. Researchers should be more careful regarding these biases when
326 conducting analyses for behavioural and other modifiable traits using biobank data sets with self-
327 reported records.

328

329 **Methods**

330 **Phenotypic data and quality controls**

331 We obtained behavioural and disease traits from the UK Biobank (UKB) data⁵⁰. UK Biobank has
332 approval from the North West Multicentre Research Ethics Committee (MREC), and informed
333 consent has been obtained from all participants. There were 455,607 individuals of European ancestry
334 with complete information on sex, age and principal components (PCs). The self-reported drinking
335 status classes (data-field ID: 20117) were: never drinkers ($n = 14,488$), previous drinkers ($n =$
336 $15,912$), current drinkers ($n = 424,507$), and unknown (446 participants preferred not to answer and
337 254 provided no response). We removed “former drinkers” from all analyses in this study,
338 considering the occurrence of the “sick quitter phenomenon”⁵¹. Among the 424,507 current drinkers,
339 there were 358,449 individuals who reported their intake level. The AC level was summed up as a
340 weekly total intake score (units/week) of all the alcoholic drink subtypes including beer plus cider, red
341 wine, champagne plus white wine, spirits, and fortified wine. The mean of AC was 10.67 units per
342 week ($s.d. = 10.23$). One unit was defined as one measure for spirits, one glass for red wine/white
343 wine/champagne, or one pint of beer/cider. The raw AC units were transformed by $\log_2(\text{raw AC units}$
344 $+ 1)$ to avoid having a heavily skewed distribution. Smoking intensity was measured in cigarettes per
345 day (CPD) in all current smokers (data-field ID: 3456; $n = 32,801$). Physical activity traits in the
346 UKB were collected from both self-reported questionnaires and devices (wrist-worn accelerometers).
347 METT is a total score of the Metabolic Equivalent Task (MET) minutes per week for walking,

348 moderate activity, and vigorous activity (data-field IDs: 864, 874, 884, 894, 904, and 914; $n =$
349 417,938). IPAQ is a derived categorical trait that utilizes the information from the METT and its
350 three subsets mentioned above (see transformation criteria via the link in the **URLs**). The three IPAQ
351 categories are denoted as low, moderate, and high ($n = 100,611, 190,056, \text{ and } 127,271$). OAA
352 (overall acceleration average) is an objective assessment of physical activity using a wrist-worn
353 accelerometer (data-field ID: 90012; $n = 97,006$). The participants were voluntary, and the
354 measurements were collected for seven days (see Doherty et al.⁵² for more details).

355

356 We extracted the phenotypic data of common diseases from the UKB following the disease
357 definitions used in Zhu et al.⁴³. There were 22 common diseases in total, and we filtered out 4 diseases
358 with a prevalence $< 3\%$ in the UKB. The mean disease count was 1.45 ($s.d. = 1.56$) in the whole
359 sample and 1.41 ($s.d. = 1.53$) in current drinkers. Body mass index (BMI) was obtained from the
360 physical measurements (data-field: 21001). Educational attainment was indexed by years of school
361 derived from qualification data (data-field ID: 6138). For quantitative traits, extreme phenotypic
362 values outside the mean $\pm 7 s.d.$ range in each sex group were excluded (see below for a more
363 stringent QC step to remove phenotypic outliers after standardisation).

364

365 **Correcting for misreports and longitudinal changes**

366 Our MLC corrections consist of two steps. The first step is a phenotypic quality control (QC)
367 procedure used as an attempt to minimize the effects of misreports. We removed the individuals who
368 self-reported as 1) never drinkers ($n = 14,488$), 2) current drinkers with a reported weekly
369 consumption of zero ($n = 9,064$), and 3) current drinkers who provided no response to AC ($n =$
370 66,058), and retained a total of 349,385 individuals. The second step is to account for self-reported
371 longitudinal changes compared to 10 years ago (data-field ID: 1628). We partitioned the individuals
372 who passed the QC above into three groups based on the longitudinal change (i.e., LESS, SAME or
373 MORE) and conducted GWAS within each group. In the LESS group, we further removed individuals
374 who reduced their AC because of being ill or doctor's advice, i.e., longitudinal change due to disease
375 ascertainment (data-field ID: 2664; $n = 11,886$). We then performed an inverse-variance weighted
376 meta-analysis of the GWAS results from the three groups ($n = 336,469$). This partitioning strategy
377 efficiently removed any difference in mean or variance between the three groups. More details of the
378 MLC corrections are shown in **Supplementary Figures 1 and 2**.

379

380 **Genome-wide association analysis**

381 The UKB genotype data were cleaned and imputed into the Haplotype Reference Consortium
382 (HRC)⁵³ panel by the UKB team²³. We selected a subset of the sample of European ancestry ($n =$
383 456,426) from the whole UKB cohort by projecting the individuals onto the PCs from the 1000

384 Genome Project (1KGP). Genotype posterior probabilities were converted to hard-call genotypes
385 using PLINK2 (--hard-call-thresh 0.1)⁵⁴. We removed SNPs with a minor allele count < 5, Hardy-
386 Weinberg equilibrium test P -value < 1×10^{-6} , missing genotype rate > 5%, or imputation info score <
387 0.3. For binary traits, we performed BOLT-LMM analysis⁵⁵ with sex, age and the first 10 PCs fitted
388 as covariates and then transformed the estimates of SNP effects on the observed 0-1 scale to odds
389 ratios (OR) by LMOR⁵⁶. For quantitative traits, we adjusted the phenotypic values for sex and age,
390 standardised the adjusted phenotypes to z-scores, excluded individuals with $|z| > 5$, and conducted
391 the BOLT-LMM analysis⁵⁵ with the first 10 PCs as fitted as covariates in the model.

392

393 Considering a loss of power due to decreased sample size by MLC corrections, we randomly down-
394 sampled the GWAS data by 21,940 individuals and repeated this process 30 times. We used a z-
395 statistic to test if the number of loci that became non-significant (or changed from non-significant to
396 significant) after the MLC corrections is significantly different from that expected by random down-
397 sampling. The average number of loci that became non-significant due to down-sampling was 10.03
398 (standard error of the mean (*s. e. m.*) = 0.85), significantly ($P = 2.08 \times 10^{-12}$) smaller than the
399 decrease in the number of genome-wide significant loci due to the MLC corrections (i.e., 16). For the
400 GWAS signals that lost due to down-sampling, the average proportion of significant associations with
401 the metabolic/cardiovascular traits in PheWAS was 31.2% (*s. e. m.* = 3.8%), which was significantly
402 lower than the observed 44.9% ($P = 3.61 \times 10^{-4}$), supporting the enrichment of the 16 loci in
403 metabolic/cardiovascular traits (**Supplementary Table 3**). We also identified 10 loci that became
404 genome-wide significant only after the MLC corrections (**Supplementary Table 2**). The down-
405 sampling analysis showed that only 3.27 loci (*s. e. m.* = 0.52) would be expected by chance
406 (**Supplementary Table 3**), indicating that most of the 10 loci were likely to be masked by MLC in
407 the uncorrected GWAS.

408

409 **Estimating heritability and genetic correlation**

410 We used the LD score regression⁵⁷ (LDSC) to estimate SNP-based heritability for a trait and the
411 bivariate-LDSC³⁵ to estimate genetic correlation between traits using ~1.2 million SNPs in common
412 with those in HapMap 3 (Ref⁵⁸). For the 234 traits for which we obtained GWAS summary data from
413 LD Hub (**URLs**), the LDSC analyses was performed online in LD Hub³⁶. Note that due to the
414 restricted access to the full summary statistics of the 23andMe data sets, we did not perform the
415 genetic correlation analysis for AC using the full GSCAN data⁵.

416

417 **Mendelian randomization analysis**

418 Mendelian randomization (MR) is a method to estimate causative effect of an exposure on an
419 outcome using instrumental variables (IVs) associated with the exposure^{39,40}. MR assumes that the
420 IVs are independent of possible confounders that may associate with both the exposure and outcome.

421 Also, the IVs are assumed not to be associated with the outcome other than mediated through the
422 exposure. However, in real data, these assumptions can be violated, leading to a biased estimate of the
423 causal effect⁵⁹. We performed MR analyses to test the causal effect of AC on BMI using IVW, MR-
424 Egger and weighted median implemented in the R package ‘MendelianRandomization’ (**URLs**), MR-
425 PRESSO in R and GSMR implemented in GCTA (v1.91.8beta) (**URLs**). The IVs were selected from
426 a clumping analysis of the GWAS summary statistics in GCTA-GSMR (clumping criteria: window
427 size = 1 Mb, $P = 5 \times 10^{-8}$ and LD $r^2 = 0.01$).

428

429 **Simulating data with disease ascertainment**

430 We carried out simulations to mimic the bias due to disease ascertainment in GWAS and its follow-up
431 analyses. If individuals who are affected by a disease tend to change a behaviour, such a change of
432 behaviour would lead to a spurious correlation between the disease and behaviour. We considered
433 four scenarios in the simulation: I) the disease liability (D) is independent of the behavioural trait (Y),
434 and 100 SNPs are associated with Y only; II) Y had a causal effect on D, and 100 SNPs are associated
435 with Y (and D mediated through Y); III) Y and D are independent, and 100 SNPs are associated with
436 D only; IV) Y had a causal effect on D, 100 SNPs affected Y (and D mediated through Y), and
437 another set of 100 SNPs affected D directly. In each scenario, to mimic the disease ascertainment, we
438 reduced the values of Y for the individuals who had high values of D. More specifically, if the D
439 value of an individual passed a threshold (e.g., top 10%), the corresponding Y value would be
440 subtracted by a constant. We set the disease ascertainment threshold to be 10%, 20%, 30%, or 40%
441 and considered a subtraction from Y value by 1, 2, 3, 4 or 5 standard deviations. We then conducted
442 GWAS and estimated SNP effect correlation between Y and D using the r_b approach developed in a
443 previous study⁶⁰ as well as the causal effect of Y on D using GSMR. To demonstrate the effectiveness
444 of our MLC corrections, we performed an adjustment for longitudinal change in our simulation
445 setting under scenario IV. We divided the individuals into two groups (LESS and SAME) and then
446 conducted the GWAS separately, followed by a meta-analysis. Details of the simulation process and
447 parameter specifications can be found in the **Supplementary Note 2**.

448

449 **URLs**

450 PheWAS: <http://atlas.ctglab.nl/PheWAS>

451 LDSC: <https://github.com/bulik/ldsc>

452 LD Hub: <http://ldsc.broadinstitute.org/ldhub/>

453 GSMR: <http://cnsgenomics.com/software/gsmr/>

454 MR-PRESSO: <https://github.com/rondolab/MR-PRESSO>

455 Other MR methods: <https://CRAN.R-project.org/package=MendelianRandomization>

456 IPAQ data processing: http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/ipaq_analysis.pdf

457

458 **Data availability**

459 The individual-level genotype and phenotype data are available through formal
460 application to the UK Biobank (<http://www.ukbiobank.ac.uk>).

461

462 **Acknowledgements**

463 We thank Kyoko Watanabe for the assistance in the PheWAS analysis. This research was supported
464 by the Australian Research Council (FT180100186 and FL180100072), and the Australian National
465 Health and Medical Research Council (1107258, 1173790, and 1113400). This study makes use of
466 data from the UK Biobank (project ID: 12505). A full list of acknowledgements can be found in the

467 **Supplementary Note 5.**

468

469 **Author contributions**

470 JY and AX conceived the study. JY, AX and JZ designed the experiment. AX performed all the
471 analyses and simulations. LJ contributed to the analysis of the physical activity data. ZZ contributed
472 to the GSMR analysis. PMV and NRW provided critical advice in data analysis and interpretation of
473 the results. AX, JY and JZ wrote the manuscript with the participation of all authors. All the authors
474 approved the final version of the manuscript.

475

476 **Competing interests**

477 The authors declare no competing financial interests.

478 **References**

- 479 1. Holmes, M.V. *et al.* Association between alcohol and cardiovascular disease: Mendelian
480 randomisation analysis based on individual participant data. *Bmj-British Medical Journal*
481 **349**(2014).
- 482 2. Griswold, M.G. *et al.* Alcohol use and burden for 195 countries and territories, 1990–2016: a
483 systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **392**, 1015-1035
484 (2018).
- 485 3. Burton, R. & Sheron, N. No level of alcohol consumption improves health. *Lancet* **392**, 987-
486 988 (2018).
- 487 4. Clarke, T.K. *et al.* Genome-wide association study of alcohol consumption and genetic
488 overlap with other health-related traits in UK Biobank (N=112 117). *Mol Psychiatry* **22**,
489 1376-1384 (2017).
- 490 5. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the
491 genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237-244 (2019).
- 492 6. Sanchez-Roige, S. *et al.* Genome-wide association study of alcohol use disorder
493 identification test (AUDIT) scores in 20 328 research participants of European ancestry.
494 *Addiction biology* (2017).
- 495 7. Sanchez-Roige, S. *et al.* Genome-wide association study meta-analysis of the Alcohol Use
496 Disorders Identification Test (AUDIT) in two population-based cohorts. *American Journal of*
497 *Psychiatry*, appi. ajp. 2018.18040369 (2018).
- 498 8. Knott, C., Bell, S. & Britton, A. Alcohol Consumption and the Risk of Type 2 Diabetes: A
499 Systematic Review and Dose-Response Meta-analysis of More Than 1.9 Million Individuals
500 From 38 Observational Studies. *Diabetes Care* **38**, 1804-1812 (2015).
- 501 9. Bell, S. *et al.* Association between clinically recorded alcohol consumption and initial
502 presentation of 12 cardiovascular diseases: population based cohort study using linked health
503 records. *Bmj-British Medical Journal* **356**(2017).
- 504 10. de Timary, P. *et al.* The loss of metabolic control on alcohol drinking in heavy drinking
505 alcohol-dependent subjects. *PLoS One* **7**, e38682 (2012).
- 506 11. Lieber, C.S. Relationships between nutrition, alcohol use, and liver disease. *Alcohol Research*
507 *& Health* **27**, 220-231 (2003).
- 508 12. Kranzler, H.R. *et al.* Genome-wide association study of alcohol consumption and use disorder
509 in 274,424 individuals from multiple populations (vol 10, 1499, 2019). *Nature*
510 *Communications* **10**(2019).
- 511 13. Holmes, M.V. & Davey Smith, G. Problems in interpreting and using GWAS of conditional
512 phenotypes illustrated by 'alcohol GWAS'. *Mol Psychiatry* (2018).
- 513 14. Munafò, M.R., Tilling, K., Taylor, A.E., Evans, D.M. & Smith, G.D. Collider scope: when
514 selection bias can substantially influence observed associations. *International Journal of*
515 *Epidemiology* **47**, 226-235 (2018).
- 516 15. Duffy, J.C. & Waterton, J.J. Under-reporting of alcohol consumption in sample surveys: the
517 effect of computer interviewing in fieldwork. *Br J Addict* **79**, 303-308 (1984).
- 518 16. Perez-Stable, E.J., Marin, B.V., Marin, G., Brody, D.J. & Benowitz, N.L. Apparent
519 underreporting of cigarette consumption among Mexican American smokers. *Am J Public*
520 *Health* **80**, 1057-1061 (1990).
- 521 17. Stockwell, T. *et al.* Under-reporting of alcohol consumption in household surveys: a
522 comparison of quantity-frequency, graduated-frequency and recent recall. *Addiction* **99**, 1024-
523 1033 (2004).
- 524 18. Hazell, L. & Shakir, S.A. Under-reporting of adverse drug reactions : a systematic review.
525 *Drug Saf* **29**, 385-396 (2006).
- 526 19. Wood, A.M. *et al.* Risk thresholds for alcohol consumption: combined analysis of individual-
527 participant data for 599 912 current drinkers in 83 prospective studies. *Lancet* **391**, 1513-1523
528 (2018).
- 529 20. Kerr, W.C., Fillmore, K.M. & Bostrom, A. Stability of alcohol consumption over time:
530 Evidence from three longitudinal surveys from the United States. *Journal of Studies on*
531 *Alcohol* **63**, 325-333 (2002).

- 532 21. Paavola, M., Vartiainen, E. & Haukkala, A. Smoking, alcohol use, and physical activity: A
533 13-year longitudinal study ranging from adolescence into adulthood. *Journal of Adolescent*
534 *Health* **35**, 238-244 (2004).
- 535 22. Moore, A.A. *et al.* Longitudinal patterns and predictors of alcohol consumption in the United
536 States. *Am J Public Health* **95**, 458-465 (2005).
- 537 23. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*
538 **562**, 203-209 (2018).
- 539 24. Rehm, J. *et al.* Are lifetime abstainers the best control group in alcohol epidemiology? On the
540 stability and validity of reported lifetime abstention. *Am J Epidemiol* **168**, 866-871 (2008).
- 541 25. Boniface, S., Scholes, S., Shelton, N. & Connor, J. Assessment of Non-Response Bias in
542 Estimates of Alcohol Consumption: Applying the Continuum of Resistance Model in a
543 General Population Survey in England. *PLoS One* **12**, e0170892 (2017).
- 544 26. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment
545 methods. *J Multidiscip Healthc* **9**, 211-217 (2016).
- 546 27. Lavrakas, P.J. *Encyclopedia of survey research methods*, (Sage Publications, 2008).
- 547 28. Boniface, S., Kneale, J. & Shelton, N. Drinking pattern is more strongly associated with
548 under-reporting of alcohol consumption than socio-demographic factors: evidence from a
549 mixed-methods study. *BMC public health* **14**, 1297 (2014).
- 550 29. Latkin, C.A. *et al.* Social desirability response bias and other factors that may influence self-
551 reports of substance use and HIV risk behaviors: a qualitative study of drug users in Vietnam.
552 *AIDS Education and Prevention* **28**, 417-425 (2016).
- 553 30. Lemmens, P., Knibbe, R.A. & Tan, F. Weekly recall and diary estimates of alcohol
554 consumption in a general population survey. *J Stud Alcohol* **49**, 131-135 (1988).
- 555 31. Ekholm, O. Influence of the recall period on self-reported alcohol intake. *Eur J Clin Nutr* **58**,
556 60-63 (2004).
- 557 32. Klatsky, A.L., Gunderson, E.P., Kipp, H., Udaltsova, N. & Friedman, G.D. Higher prevalence
558 of systemic hypertension among moderate alcohol drinkers: an exploration of the role of
559 underreporting. *Journal of Studies on Alcohol* **67**, 421-428 (2006).
- 560 33. Lahaut, V.M., Jansen, H.A., van de Mheen, D. & Garretsen, H.F. Non-response bias in a
561 sample survey on alcohol consumption. *Alcohol Alcohol* **37**, 256-260 (2002).
- 562 34. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex
563 traits. *Nat Genet* **51**, 1339-1348 (2019).
- 564 35. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat*
565 *Genet* **47**, 1236-1241 (2015).
- 566 36. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score
567 regression that maximizes the potential of summary level GWAS data for SNP heritability
568 and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
- 569 37. Collins, S.E. Associations Between Socioeconomic Factors and Alcohol Outcomes. *Alcohol*
570 *Research-Current Reviews* **38**, 83-94 (2016).
- 571 38. Davies, N.M., Dickson, M., Davey Smith, G., van den Berg, G.J. & Windmeijer, F. The
572 Causal Effects of Education on Health Outcomes in the UK Biobank. *Nat Hum Behav* **2**, 117-
573 125 (2018).
- 574 39. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute
575 to understanding environmental determinants of disease? *International Journal of*
576 *Epidemiology* **32**, 1-22 (2003).
- 577 40. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N. & Smith, G.D. Mendelian
578 randomization: Using genes as instruments for making causal inferences in epidemiology.
579 *Statistics in Medicine* **27**, 1133-1163 (2008).
- 580 41. Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian Randomization Analysis With
581 Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* **37**, 658-665
582 (2013).
- 583 42. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid
584 instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*
585 **44**, 512-525 (2015).

- 586 43. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from
587 GWAS summary data. *Nature communications* **9**, 224 (2018).
- 588 44. Bowden, J., Smith, G.D., Haycock, P.C. & Burgess, S. Consistent Estimation in Mendelian
589 Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic*
590 *Epidemiology* **40**, 304-314 (2016).
- 591 45. Verbanck, M., Chen, C.Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy
592 in causal relationships inferred from Mendelian randomization between complex traits and
593 diseases. *Nat Genet* **50**, 693-698 (2018).
- 594 46. Millwood, I. *et al.* Conventional and genetic evidence on alcohol and vascular disease
595 aetiology: prospective study of 500,000 Chinese adults. *Lancet* (2018).
- 596 47. Sabia, S. *et al.* Association Between Questionnaire- and Accelerometer-Assessed Physical
597 Activity: The Role of Sociodemographic Factors. *Am J Epidemiol* **179**, 781-790 (2014).
- 598 48. Knott, C.S., Coombs, N., Stamatakis, E. & Biddulph, J.P. All cause mortality and the case for
599 age specific alcohol consumption guidelines: pooled analyses of up to 10 population based
600 cohorts. *Bmj-British Medical Journal* **350**(2015).
- 601 49. Evangelou, E. *et al.* New alcohol-related genes suggest shared genetic mechanisms with
602 neuropsychiatric disorders. *Nat Hum Behav* (2019).
- 603 50. Bycroft, C. *et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. *bioRxiv*,
604 166298 (2017).
- 605 51. Klatsky, A.L. & Udaltsova, N. Abounding confounding: sick quitters and healthy drinkers.
606 *Addiction* **108**, 1549-1552 (2013).
- 607 52. Doherty, A. *et al.* Large Scale Population Assessment of Physical Activity Using Wrist Worn
608 Accelerometers: The UK Biobank Study. *PLoS One* **12**, e0169649 (2017).
- 609 53. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
610 *Genet* **48**, 1279-1283 (2016).
- 611 54. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
612 datasets. *Gigascience* **4**, 7 (2015).
- 613 55. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large
614 cohorts. *Nat Genet* **47**, 284-290 (2015).
- 615 56. Lloyd-Jones, L.R., Robinson, M.R., Yang, J. & Visscher, P.M. Transformation of Summary
616 Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio.
617 *Genetics, genetics*. 300360.302017 (2018).
- 618 57. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity
619 in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015).
- 620 58. International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse
621 human populations. *Nature* **467**, 52-58 (2010).
- 622 59. Evans, D.M. & Davey Smith, G. Mendelian Randomization: New Applications in the Coming
623 Age of Hypothesis-Free Causality. *Annu Rev Genomics Hum Genet* **16**, 327-350 (2015).
- 624 60. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and
625 methylomic data from blood. *Nat Commun* **9**, 2282 (2018).
- 626

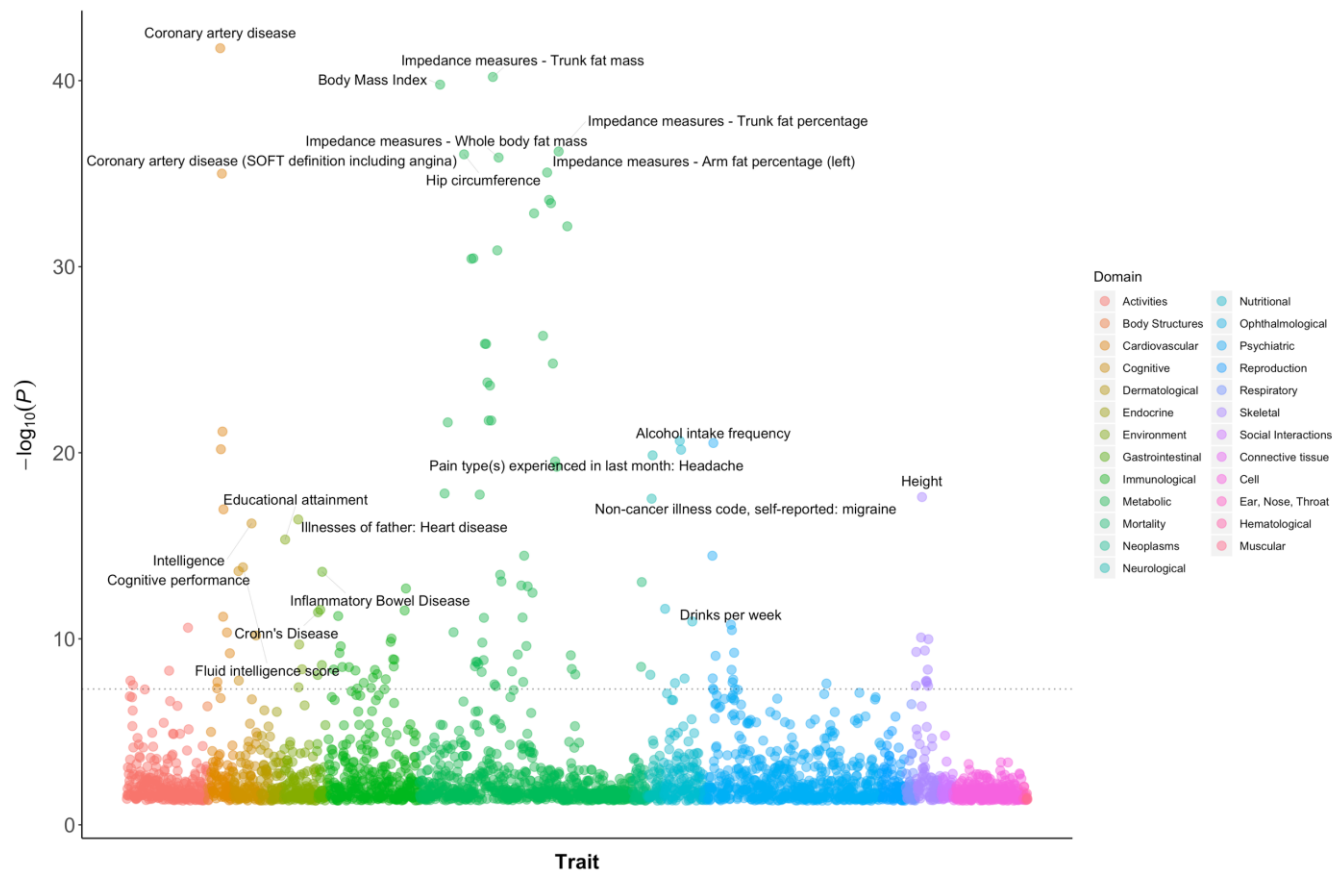


Figure 1. PheWAS results for the 16 AC GWAS signals that became non-significant after the MLC corrections. The colour denotes the domain of the associated trait. There were 136 traits associated with the 16 SNPs with $P < 5 \times 10^{-8}$, and 61 (44.9%) of them were metabolic/cardiovascular traits.



Figure 2. Estimates of genetic correlation between AC and common diseases in the UKB. The rows denote 8 GWAS summary data sets for AC with the sample size labelled in the bracket. The columns are 18 common diseases and disease count. The nominal significant effects ($P < 0.05$) are labelled with \hat{r}_g [95% confidence interval] (P -value), and the significant effects after multiple testing correction ($P < 0.05/152$) are labelled with an additional asterisk. “Current drinkers excluding underreporting” represents current drinkers excluding 9,064 individuals who likely underreported their AC levels. LESS, SAME, and MORE represent current drinkers whose AC levels were reduced, maintained the same, and increased, respectively, compared to 10 years ago. “LESS with illness removed” represents the LESS group excluding the participants who reduced their AC intake level due to illness or doctor’s advice.

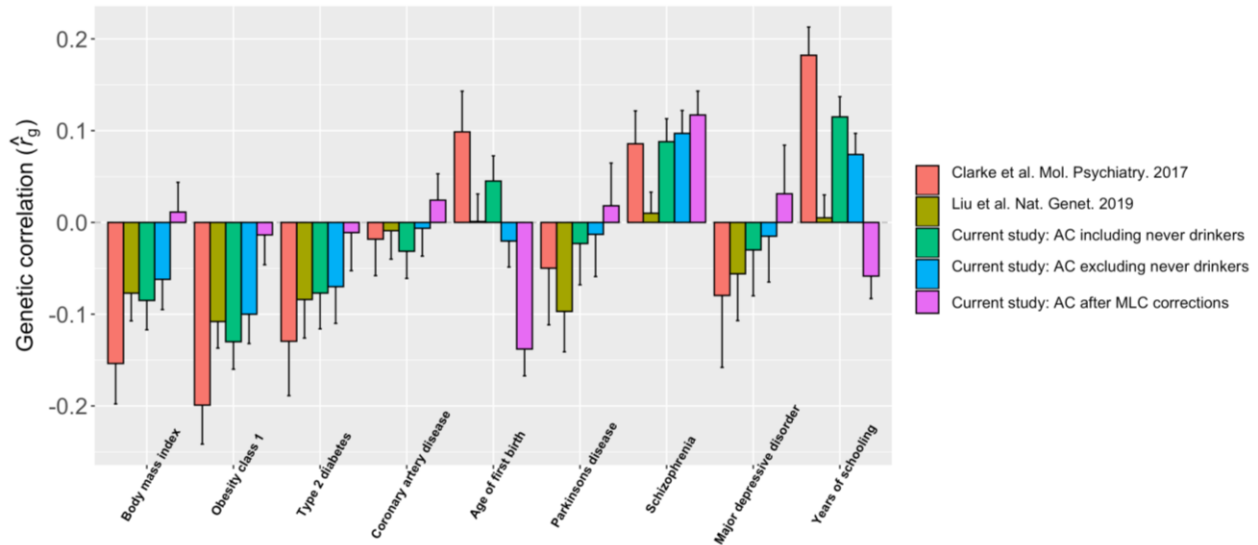


Figure 3. Estimates of genetic correlation between AC and complex traits using data from the UKB and other published studies. Genetic correlation was estimated by the bivariate-LDSC in LD Hub. The y-axis shows the estimate of r_g , and the x-axis shows different complex traits. The error bars denote the standard errors of the estimates. The results using the summary statistics from our analysis were compared to those from Clarke et al. (2017, *Molecular Psychiatry*)⁴, who used self-reported AC from the interim release of the UKB data, and Liu et al. (2019, *Nature Genetics*)⁵, a meta-analysis that included the full release of the UKB data.

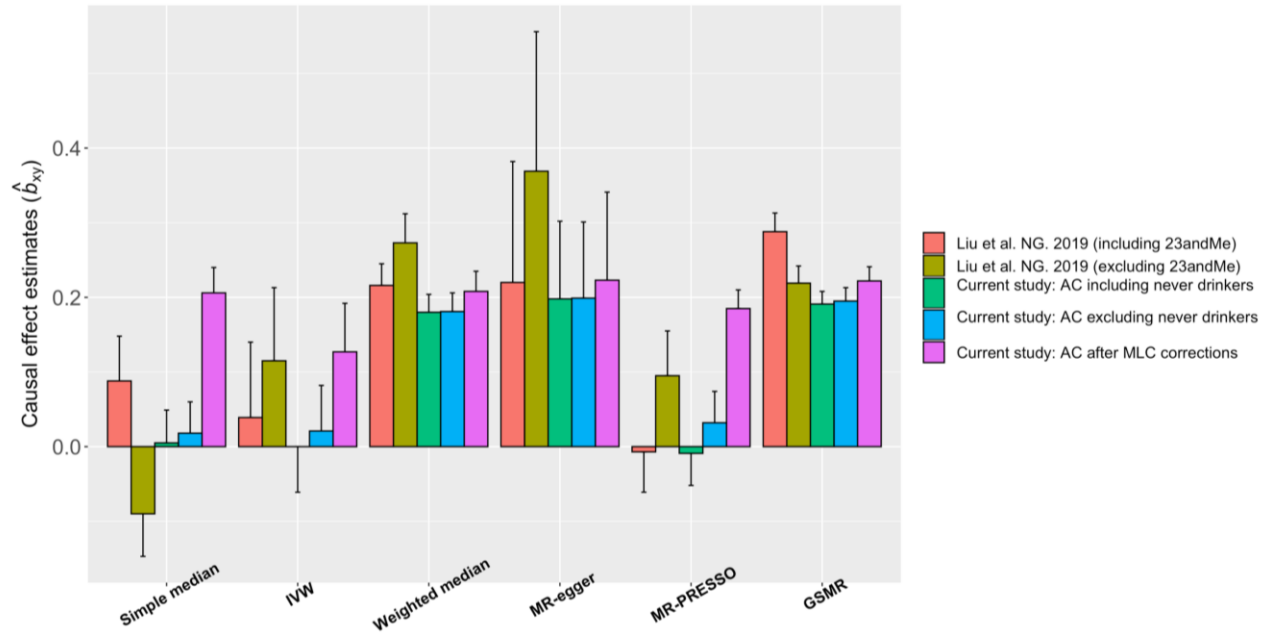


Figure 4. Estimates of causal effect of AC on BMI using different MR methods. The colour denotes five different GWAS summary data sets for AC. The y-axis shows the estimate of causal effects, and the x-axis shows different MR methods. The error bars denote the standard errors of the estimates.

Table 1. Alcohol consumption and health-related traits of current drinkers in different response and longitudinal change groups

Intake level reported?	Longitudinal change group	N	Mean phenotype			Disease prevalence			
			AC	BMI	Disease count	CVD	T2D	DYSLIPID	HYPER
Yes	All	358,449	10.67	27.17	1.372	0.154	0.051	0.169	0.184
	LESS	152,854	8.56	27.77	1.506	0.178	0.072	0.192	0.211
	SAME	134,808	10.82	26.72	1.274	0.141	0.038	0.158	0.167
	MORE	68,855	15.11	26.68	1.266	0.128	0.031	0.141	0.154
No	All	66,058	/	28.28	1.630	0.177	0.082	0.181	0.213
	LESS	38,799	/	28.70	1.730	0.194	0.096	0.195	0.229
	SAME	25,055	/	27.70	1.474	0.153	0.061	0.159	0.189
	MORE	1,599	/	27.40	1.448	0.138	0.049	0.147	0.162

LESS, SAME, and MORE denote the current drinkers who reduced, maintained, and increased their alcohol consumption, respectively, compared to 10 years ago. N, sample size; AC, alcohol consumption measured by units per week; BMI, body mass index; Disease count, number of common diseases affected; CVD: cardiovascular disease; T2D: type 2 diabetes; DYSLIPID: dyslipidaemia; and HYPER: hypertensive disease. The proportion of the LESS group was 42.6% in the current drinkers who reported the intake level and 58.7% in the current drinkers who did not report the intake level.

Table 2. Descriptive statistics of the reasons for reducing alcohol intake

Intake level reported?	Reason of reducing	N	Mean phenotype			
			AC	BMI	Disease count	EA
Yes	Illness or ill health	8,555	7.33	28.53	2.77	12.81
	Doctor's advice	3,331	14.72	29.49	2.58	12.65
	Health precaution	48,483	9.98	27.44	1.54	14.17
	Financial reasons	7,323	10.27	28.54	1.74	11.57
	Other reason	68,066	7.54	27.71	1.28	13.90
	Prefer not to answer	268	5.77	27.69	1.75	10.90
	Do not know	16,729	7.28	27.96	1.35	13.00
No	Illness or ill health	3,541	/	29.28	2.81	12.35
	Doctor's advice	462	/	29.99	2.77	11.90
	Health precaution	6,767	/	28.26	1.87	12.86
	Financial reasons	1,600	/	29.05	1.91	11.26
	Other reason	21,362	/	28.62	1.51	12.88
	Prefer not to answer	101	/	28.23	2.10	9.56
	Do not know	4,934	/	28.79	1.53	11.89

N, sample size; AC, alcohol consumption (units per week); BMI, body mass index; Disease count, number of common diseases affected; EA, educational attainment (years of schooling). A total of 132 individuals did not have records for their reasons for reducing their alcohol intake.