PheWAS-based clustering of Mendelian Randomisation instruments reveals distinct mechanism-specific causal effects between obesity and educational attainment

Liza Darrous[1,2,3,†], Gibran Hemani[4,5], George Davey Smith[4,5], and Zoltán Kutalik[1,2,3,†]

[1]University Center for Primary Care and Public Health, University of Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[4]Medical Research Council Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, United Kingdom
[5]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[†]*Correspondence should be addressed to darrous.liza@gmail.com or zoltan.kutalik@unil.ch*

Abstract

Mendelian Randomisation (MR) is a statistical method that estimates causal effects between risk factors and common complex diseases using genetic instruments. Heritable confounders, pleiotropy and heterogeneous causal effects violate MR assumptions and can lead to biases. To tackle these, we propose an approach employing a PheWAS-based clustering of the MR instruments (PWC-MR). We apply this method to revisit the surprisingly large apparent causal effect of body mass index (BMI) on educational attainment (EDU): $\widehat{\alpha}$ = -0.19 [-0.22, -0.16].

As a first step of PWC-MR, we clustered 324 BMI-associated genetic instruments based on their association profile across 407 traits in the UK Biobank, which yielded six distinct groups. The subsequent cluster-specific MR revealed heterogeneous causal effect estimates on EDU. A cluster strongly enriched for traits related to socio-economic position yielded the largest BMI-on-EDU causal effect estimate ($\widehat{\alpha}$ = -0.49 [-0.56, -0.42]) whereas a cluster enriched for primary impact on body-mass had the smallest estimate ($\widehat{\alpha}$ = -0.09 [-0.13, -0.05]). Several follow-up analyses confirmed these findings: (i) within-sibling MR results ($\widehat{\alpha}$ = -0.05 [-0.09, -0.01]); (ii) MR for childhood BMI on EDU ($\widehat{\alpha}$ = -0.03 [-0.06, -0.002]); (iii) step-wise multivariable MR (MVMR) ($\widehat{\alpha}$ = -0.06 [-0.09, -0.04]) where time spent watching television and past tobacco smoking (two proxies for potential confounders) were jointly modelled.

Through a detailed examination of the BMI-EDU causal relationship we demonstrated the utility of our PWC-MR approach in revealing distinct pleiotropic pathways and confounder mechanisms.

1 Introduction

Genome-wide association studies[1] (GWASs) have identified many genetic variants associated with multiple complex phenotypes, aiding us in annotating single nucleotide polymorphisms (SNPs) and their functions, as well as identifying putative causal genes. As sample sizes of GWASs increase, more SNP associations are revealed which improve various downstream analyses such as polygenic score prediction, pathway- and tissue-enrichment, and causal inference[2,3].

Mendelian Randomisation[4,5] (MR), an approach generally applied through the use of genetic variants/SNPs as instrumental variables (IVs) to infer the causal relationship between an exposure or a risk factor X and an outcome Y of interest, has become increasingly used thanks to well-powered GWASs from which hundreds of genetic associations with heritable exposures can be used as IVs.
MR has three major assumptions concerning the the genetic variant G used as an instrument: (1) Relevance – G is strongly associated with the exposure. (2) Exchangeability – there is no confounder of the G-outcome relationship. (3) Exclusion restriction – G affects the outcome only through the exposure. Each instrument provides a causal effect estimate, which can then be combined with others using an inverse variance-weighting[6] (IVW) method to obtain an estimate of the total causal effect of the exposure on the outcome. This estimate is more reliable than observational associations due to it being more protected against unmeasured confounding and reverse causality, provided that the core conditions are met.

Thanks to well-powered GWASs we have also discovered that most genetic instruments are highly pleiotropic[7], i.e. associated to more than a single trait. This has also been shown in phenome-wide association studies (PheWASs), where associations between a SNP and a large number of phenotypes are tested. The situation when a genetic variant influences multiple traits, but there is a primarily associated trait and all other trait associations are fully mediated by the primary trait, is referred to as vertical pleiotropy. On the other hand, genetic variants that affect some traits through pathways other than the primary trait (the exposure) – a phenomena known as horizontal pleiotropy – are in direct violation of the exclusion restriction assumption and could lead to biased causal effect estimates. However, if the InSIDE assumption [8](Instrument Strength is Independent of the Direct Effect on the outcome) holds and the direct SNP effects are on average null, then IVW will yield consistent causal effect estimates. There have been MR extensions to IVW such as MR-Egger to produce less biased causal effect estimates if the InSIDE assumption holds and direct effects are not null on average. Note that violation of the InSIDE assumption leads to correlated pleiotropy, which can severely bias causal effect estimates. Such phenomenon may emerge as a result of a heritable confounder of the exposure-outcome relationship and has been modelled in the past[9,10].
Well-powered GWAS may also provide confounded genetic associations through dynastic effects[3,11], assortative mating[12,13], and population stratification[14]. These phenomena can introduce correlation between an instrument and confounding factors, such as parental/partner traits or genetic ancestry leading to a violation of the exchangeability assumption and biased causal effect estimates. This type of confounding can be resolved when using family-based study designs[15,16] such as sibling-pair studies. Since genetic differences between sibling pairs are due to independent and random meiotic events, these effects are unaffected by population stratification and other potential confounders influencing the phenotype. This and other family-based designs have been used to obtain unbiased heritability estimates, validate GWAS results and test for unbiased causal effect estimates using MR[17,18].

Another factor that can lead to complications in MR studies is the presence of heterogeneous causal effects emerging due to distinct biological mechanisms: various subtypes of the exposure (e.g. subcutaneous vs visceral adiposity) or different biological pathways through which the

exposure impacts the outcome (e.g. interaction between the exposure and other factors). To date, confounding of genetic associations, horizontal pleiotropy and heterogeneous causal effect have been largely treated as distinct mechanisms in MR modelling. However, what they have in common is that they can lead to variable causal effects estimated depending on the group of IVs used in the MR.

To address this, we introduce in this paper our approach of PheWAS-driven clustering of instrumental variables (PWC-MR) and test the resulting clusters for distinct pathways or mechanisms that could underlie the overall causal effect of the exposure. Throughout the paper, we demonstrate the approach through the example of estimating the causal effect of body mass index (BMI) on educational attainment (EDU). This relationship has been analysed extensively in the past and family studies have shown that an apparent strong effect of higher BMI on lower educational attainment is shrunk to near zero when using family studies[17]. One explanation is that offspring BMI is influenced by parental alleles associated with parental (rearing) behaviour, which in turn modify the environment of the offspring. Such parental traits act as a confounder of the offspring genotype-EDU relationship, hence violate the exchangeability assumption of MR. Moreover, they confound the BMI-EDU association in the tested sample, violating the exclusion-restriction assumption and inducing correlated pleiotropy (see Figure 1**a**). Thus, it is plausible that some of the detected IV clusters arise through parental genetic confounding which may manifest statistically as horizontal pleiotropy. To test this, we ran a systematic confounder search and probed the causal effect of the exposure conditional on candidate confounder traits.

2 Methods

2.1 Instrumental variable selection and PheWAS

As our primary analysis, we aimed to investigate the potential pleiotropy-patterns emerging from the grouping of IVs that are strongly associated with an exposure of interest, as outlined in Figure 1**b**. With BMI selected as the exposure trait, we obtained IVs from the Neale group's UK Biobank GWAS analysis[19] (data sources can be found in Supplementary Table 1) by filtering for genome-wide significant SNPs (i.e. association p-value less than 5×10^{-8}) followed by linkage disequilibrium (LD)-based clumping using the TwoSampleMR R package[20] with the following parameters: $clump\_kb = 10,000$, $clump\_r2 = 0.001$, $pop =$ "EUR" to obtain independent IVs.

This left us with 348 BMI-associated IVs, for which we ran PheWASs with $1,480$ traits from the Neale group UK Biobank GWAS analysis[19]. We extracted for each trait and for each SNP the association effect and the corresponding standard error, creating a data matrix of 348 SNPs by $1,480$ traits. For the $1,480$ traits, we also extracted details such as variable type, origin and complete sample size, among others.

2.1.1 Quality control

We removed traits from the PheWAS data matrix that had missing association effects as well as duplicates (keeping the most recent version). Furthermore, we filtered out traits for which the effective sample size was less than $50,000$ due to their low power of association, leaving us with 424 traits.
Using genetic correlation data from the Neale group[19], we further removed traits that had a high genetic correlation with BMI, i.e. the exposure, ($r^g > 0.75$), to avoid obvious repetitions of traits closely related to it. The resulting association effect data matrix of 348 SNPs and 407 traits was then standardised (SNP effects are on a SD/SD scale) and used for further analysis. Note that for simplicity, effect sizes for binary traits were treated as those of continuous traits.

In order to test for invalid IVs, we performed a trait-wide variant of Steiger-filtering[21]. Specifically, for each SNP, we tested if any of the traits had a significantly stronger (in terms of explained variance) association compared to that of the exposure. The significance threshold for this one-sided t-test was corrected for using the total number of traits remaining (p-value $< 0.05/407$). This revealed 24 SNPs more strongly associated to traits other than BMI (such as 'Whole body water mass', 'Basal metabolic rate' and 'Sitting height') that were then removed from further analysis.

2.2 K-means clustering and trait identification

With the aim of discovering distinct meaningful groups of SNPs among the 324 IVs, we proceeded with the clustering of the SNPxTrait association effect matrix using the K-means algorithm[22]. Taking the absolute standardised effects matrix, we normalised the data frame with respect to the SNPs such that the variance of the SNP effects across all the traits equalled 1. We used the absolute effects to cluster, in order to ensure that negatively correlated traits were considered similar by the Euclidean distance based similarity measure of the k-means clustering. We then compared the performance of the clustering with different number of clusters ranging from two to 50, by measuring the Akaike Information Criterion (AIC). After finding the number of clusters with the lowest AIC score (six clusters), we proceeded with the assignment of each SNP to one of the six clusters.

In order to identify traits that were particularly associated to SNPs in each of the six clusters,

we computed an enrichment ratio (ER) in this way:
For each trait t, we calculated the per-SNP average squared effect in a given cluster j, denoted as $\sigma^2_{j,t}$. Given that SNP i belongs to cluster j, $\sigma^2_{j,t}$ was calculated as follows:

$$\sigma^2_{j,t} = \frac{1}{|c_j|} \sum_{i \in c_j} \beta^2_{i,t}$$

where c_j represents the set of SNPs in cluster j and $|c_j|$ its cardinality. We then normalised these per-SNP average squared effects for each cluster (k total clusters) across all traits to obtain the enrichment ratio (ER), $R_{j,t}$:

$$R_{j,t} = \frac{\sigma^2_{j,t}}{\frac{1}{K}\sum_{k=1}^{K} \sigma^2_{k,t}}$$

where K is the total number of clusters. For each cluster (j), traits were then prioritised by the (highest) value of ER ($R_{j,t}$).

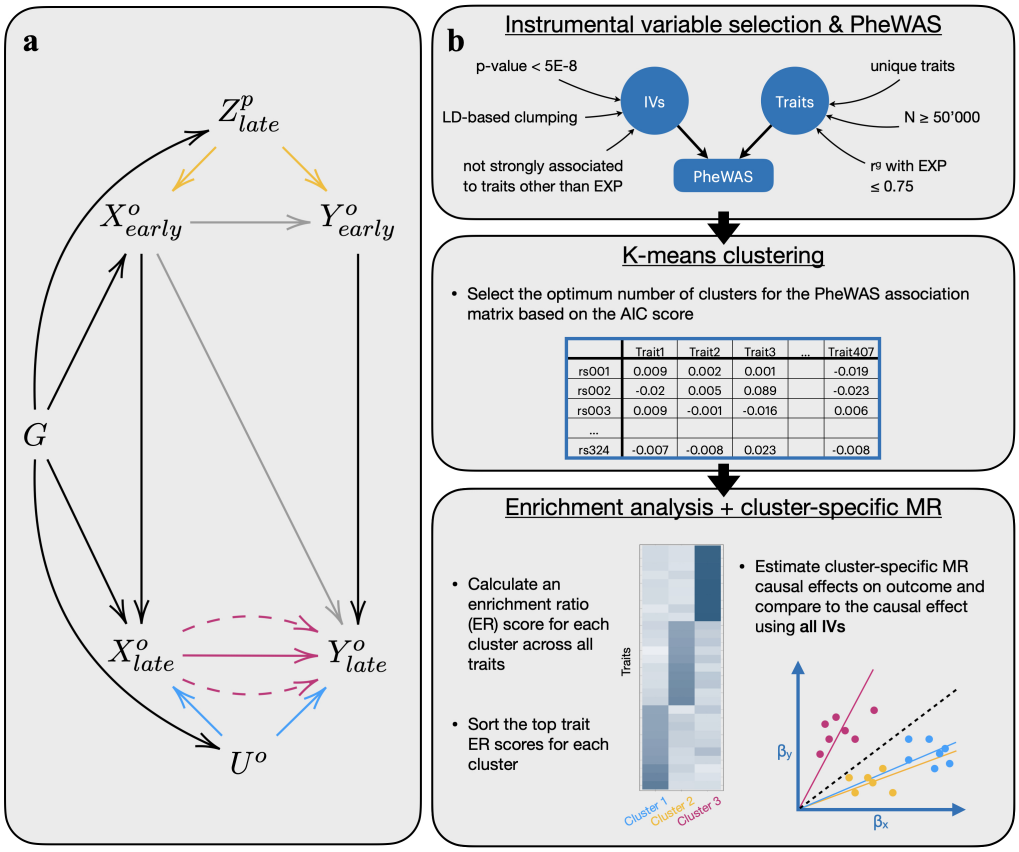

Figure 1: Directed Acyclic Graph (DAG) describing a plausible model and the flow diagram representing how the PWC-MR approach aims to disentangle causal effect between trait pairs from confounding or pleiotropy. Panel **a** illustrates a DAG involving early and later-in-life (late) versions of different traits. In our example: X and Y could be BMI and EDU, respectively, U represents a heritable confounder, whereas Z represents a parental trait involved in exerting dynastic effects. The superscripts p and o stand for parental and offspring respectively, and the dashed arrows from X to Y represent the different biological mechanisms through which a causal effect can emerge. Panel **b** represents the main steps of the PWC-MR method: (i) Instrument selection and PheWAS; (ii) IV clustering; and (iii) Enrichment analysis and cluster specific MR.

2.2.1 Causal effect estimate per cluster

We measured the cluster-specific IVW causal effect estimate on the outcome (EDU) using the standardised SNP effects in each cluster, and then compared these estimates to the causal effect estimate using all SNPs. We used the TwoSampleMR R package[20] for this analysis, and although we use two-sample MR techniques despite having a close to complete sample overlap, this does not lead to substantial biases[23]. Measures of IV heterogeneity are calculated using the Cochran's Q-statistic[24] for the IVW method for each cluster. Furthermore, average cluster-heterogeneity (per-IV variance) is also calculated for each cluster from the above-mentioned parameter.

As sensitivity analyses, PWC-MR was repeated twice, once with a different exposure trait (replacing BMI with childhood BMI), and another with a different outcome trait (replacing EDU with systolic blood pressure).

2.3 Systematic confounder search

In order to decide which of the emerging clusters represent genetic confounding or true biological heterogeneity, we systematically searched for BMI-EDU confounders. To do this, we investigated the bi-directional causal effects that each trait had on both the exposure and the chosen outcome. Firstly, an extra filtering step was done where traits that were highly genetically correlated with the outcome ($r^g > 0.75$) were removed from the total 407 traits of the previous analysis.
Then we ran a bidirectional MR for the remaining uncorrelated traits using the TwoSampleMR R package[20], and obtained four sets of causal effect measurements per trait (bidirectional, two different outcome traits - BMI and EDU). To select bidirectional causal effect estimates from those calculated by the different methods in the TwoSampleMR package[20] (Weighted median, Inverse variance weighted, Simple mode, and Weighted mode), we ordered the p-values of the causal effect estimates for the different methods and selected the estimate of the second most significant method to ensure that at least one other method supports the causal claim.

The next step was to identify the direction of causality. To do so, we performed a one-sided t-test on the estimated causal effect between the trait and the exposure, BMI. If the t-test association p-value was < 0.05, then the trait had a (nominally) significantly larger effect on the exposure, and if the p-value was ≥ 0.95, then the exposure had a (nominally) significantly larger effect on the trait. For all the p-values in between, it was challenging to assign a direction in which the causal effect was stronger, and thus these traits were not further categorised. The p-value thresholds we apply are not intended to suggest that there is a transition point at which the meaning of associations change. Rather we use these as a heuristic that is required to control type I error rate at an arbitrary (5%) threshold.
The same was done to explore the relationship between the traits and the outcome trait (EDU). This allowed us to classify the traits into candidate confounders, mediators, colliders and other categories (as seen in the middle panel of Supplementary Figure S1). For example, a confounder was defined as a trait with a significantly larger effect on both exposure and outcome than the reverse. We then focused only on the confounders which can distort MR estimates and filtered them further to make sure that they have at least a nominally significant MR estimate (p-value < 0.05) on both BMI and EDU. We were lenient in our categorisation of candidate traits as adding potentially irrelevant traits would not bias the multivariable causal effect of BMI in the next step. Mediators and colliders were not considered further since their inclusion into an MVMR does not alter the exposure's causal effect. The same holds for traits with a direct effect on either the exposure or the outcome only.
Furthermore, to test how compatible the two lines of analysis were, we examined the cluster-specific enrichment ratio values for the set of candidate confounder traits we obtained.

2.3.1 Multivariable MR

Focusing on the candidate confounder traits resulting from the systematic search that could bias the causal effect estimate between the exposure-outcome pair, we first ran a stepwise multivariable MR (MVMR) (adapted from the bGWAS R package[25]) with them as exposures to test their effect on our chosen outcome, EDU.

To do this, we created a Z-score matrix combining genome-wide significant (p-value less than $5 \times 10^{-}8$) and independent (LD-clumped $clump\_kb = 5,000$, $clump\_r2 = 0.01$) SNPs and their Z-scores for the candidate confounder traits, given that each SNP had an effect that is genome-wide significant for at least one of the candidate traits. We then removed any trait that had less than three instruments, leaving us with a Z-score matrix of 274 SNPs and 5 traits . Using this Z-score matrix as input for step-wise MVMR, we obtained a final list of traits with multivariable causal effects with a p-value $< 0.05/5$ on our chosen outcome. To ensure the strength of the instruments used for running MVMR, we calculated the conditional F-statistic for our main exposure (BMI) given each of the surviving traits and their different combinations. We then ran standard MVMR using the combination of traits with a conditional F-statistic of BMI > 8, and BMI as exposures to estimate the conditional causal estimates on EDU.

We were more lenient with the conditional F-statistic threshold (typically 10)[26] to ensure that epidemiologically relevant traits are included in the MVMR.

2.4 Relation to other approaches

2.4.1 Comparison against MR-Clust

We compared the k-means clustering of BMI IVs against another IV clustering method called MR-Clust[27], which requires as input the unstandardised SNP effects on both the exposure and the outcome, as well as the standard error of the SNP on each. To do so, we performed a Fisher's exact test to examine the frequency distribution of SNPs in each of the k-means clusters against the MR-Clust clusters.

2.4.2 Colocalisation analysis

To further interpret the findings of the IV clustering, we sought to test if specific patterns of colocalisation in different tissue types appear for the different IV clusters.

To do this, we reran the steps detailed in Leyden et al.[28] for the 324 BMI IVs used in this work. For each IV, we tested for genetic colocalisation between the BMI GWAS data and the gene expression (eQTL) data of both subcutaneous adipose and brain tissue (data sources can be found in Supplementary Table 1). For each SNP tested, we took a margin of 200kb up- and downstream, and used the coloc R package[29] to test the SNP's colocalisation with each gene found in that region, once using brain eQTL data, and another colocalisation using adipose eQTL data. We declared colocalisation if the posterior probability of the model sharing a single causal variant was larger than 80%. For each of the aforementioned clusters, we investigated if the IVs were more strongly enriched for or depleted in one tissue or the other using Fisher's exact test.

3 Results

3.1 Overview of the method

We applied the PWC-MR approach to investigate potential horizontal pleiotropic effects (emerging due to heritable confounders, dynastic effects, genetic subtypes of obesity and other pleiotropic mechanisms, see Figure 1**a**) of BMI on educational attainment. The analysis focused on grouping the IVs of the exposure by running a PheWAS-based clustering to reveal distinct mechanisms or pathways underlying their overall effect on the exposure (Figure 1**b**). This was done by obtaining the standardised PheWAS association of the BMI IVs across a filtered set of 408 traits, and running a k-means clustering on the resulting matrix. This resulted in six clusters of IVs for BMI, which were then annotated by traits based on the association of the cluster-member SNPs with each trait. Specifically, for each cluster-trait pair we computed the average explained variance of the trait by the SNPs of the given cluster. This yielded for each cluster-trait pair an enrichment ratio (ratio of the average explained variances) and we chose the top ten traits with the highest enrichment ratio for each cluster. Furthermore, the causal effect of each cluster's IVs on education was calculated and compared against each other and that of the causal effect obtained using all BMI IVs.

To complement our findings from the clustering-based analysis, we explored (i) the BMI-EDU causal relationship using sib-regression SNP effect sizes[18], (ii) the childhood BMI-EDU causal relationship, (iii) replacing the outcome trait with systolic blood pressure (SBP), and finally (iv) the potential role of each of the filtered set of traits as a confounder of the BMI-EDU relationship.
We implemented the latter one by systematically running bidirectional MR between each of the traits and either BMI or EDU as outcome, then classifying the traits depending on their bidirectional associations with both BMI and EDU. The resulting set of candidate confounder traits was further analysed for its potential to bias the causal effect of BMI on EDU. To assess this, we ran stepwise MVMR and finally calculated the causal effect of BMI on EDU conditional on the surviving set of candidate confounder traits of the BMI-EDU relationship.

To further understand the emerging clusters, we sought to uncover tissue-specific mechanisms. To do this, we performed a colocalisation analysis of the BMI and gene expression association signals at each locus around (±400kb) the 324 BMI IVs. For the gene expression association we used eQTL data from both adipose and brain tissue. This yielded a proportion of brain-vs-adipose colocalised IVs for each cluster.

3.2 PheWAS-based K-means clustering and trait identification

After identifying 324 genome-wide significant SNPs as IVs for BMI, and selecting 407 filtered traits to run PheWAS on, we obtained a standardised effect matrix of the 324 IVs on the 407 traits. Normalising the matrix by IVs and running K-means clustering on it revealed that six clusters yielded the lowest AIC score (Supplementary Figure S2) when compared to varying the number of clusters from two to 50. The number of SNPs in each of the six clusters were: $32, 98, 35, 41, 69, 49$ respectively (Supplementary Table 2).

Next, we computed an enrichment ratio (ER) to identify with which traits the SNPs in each cluster were strongly associated. The overall ER value between clusters was roughly centred around 1, however clusters #2, #3, #4, and #6 had some large ER values (see Supplementary Figure S3). Visualising the top 10 enriched traits in each cluster and their ER values in Figure 2 and Supplementary Table 3, we see that cluster #2 is strongly enriched for lean mass traits such as 'Trunk fat-free mass' and 'Whole body fat-free mass'.

Similarly, cluster #3 seemed to mostly be enriched for blood- and body stature-related traits such as 'Platelet count' and 'Standing height', while cluster #4 was enriched for traits related to socio-economic position (SEP) such as 'Job involves heavy manual or physical work', 'Time spent outdoors in summer', and 'Fluid intelligence score'. Lastly cluster #6 was enriched for food supplements/nutrients such as 'Folate' and 'Potassium'.

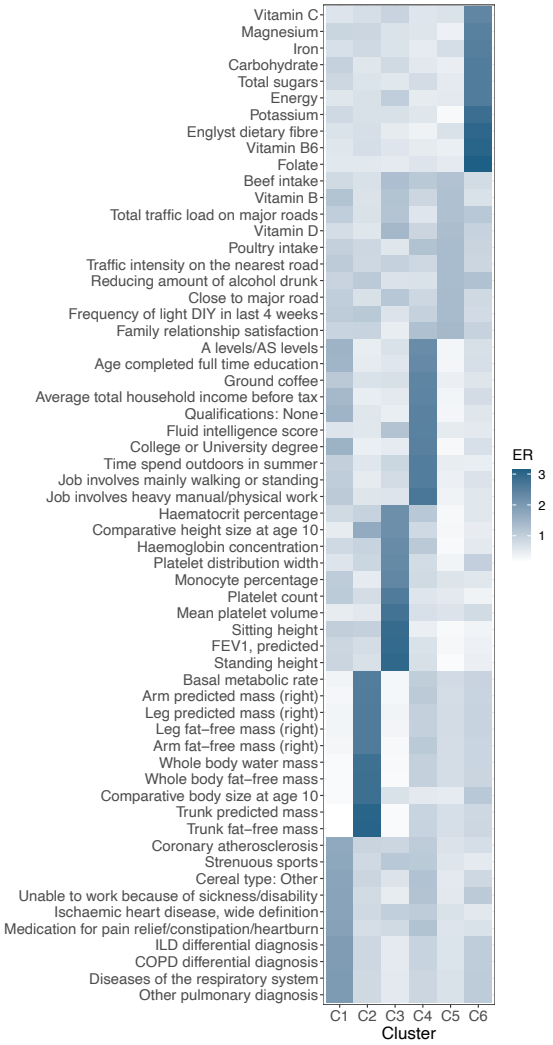

Figure 2: Heatmap of the enrichment ratio of the top 10 traits in each cluster. K-means clustering of BMI revealed six clusters with the following trait enrichment ratios.

3.2.1 Causal effect estimate per cluster

To test whether the clusters had different causal effects on a selected outcome than the overall causal effect (using all IVs), we computed the IVW causal effect estimate of each cluster on education using cluster-specific IVs. As seen in Figure 3**a** and Supplementary Table 4, the causal effect estimates between the different clusters are significantly heterogeneous (Q-test value = 130.61, p-value $< 10^{-300}$). Clusters #2 and #5 had the smallest causal effect estimates of -0.09 (p-value $= 1.23 \times 10^{-}5$) and -0.12 (p-value $= 5.22 \times 10^{-}6$) respectively, where cluster #2 was enriched for lean-mass traits. These estimates are consistent with those obtained from within-family studies, which are relatively immune to confounding. By contrast, clusters #1 and #4 had the largest negative causal effect estimates of -0.44 (p-value $= 7.78 \times 10^{-}20$) and -0.49

(p-value = $1.63 \times 10^{-}44$) respectively, where cluster #4 was strongly enriched for SEP-related traits.

All the clusters were less heterogeneous than the group of all the IVs combined (see 'Avg_het' in Supplementary Table 4).

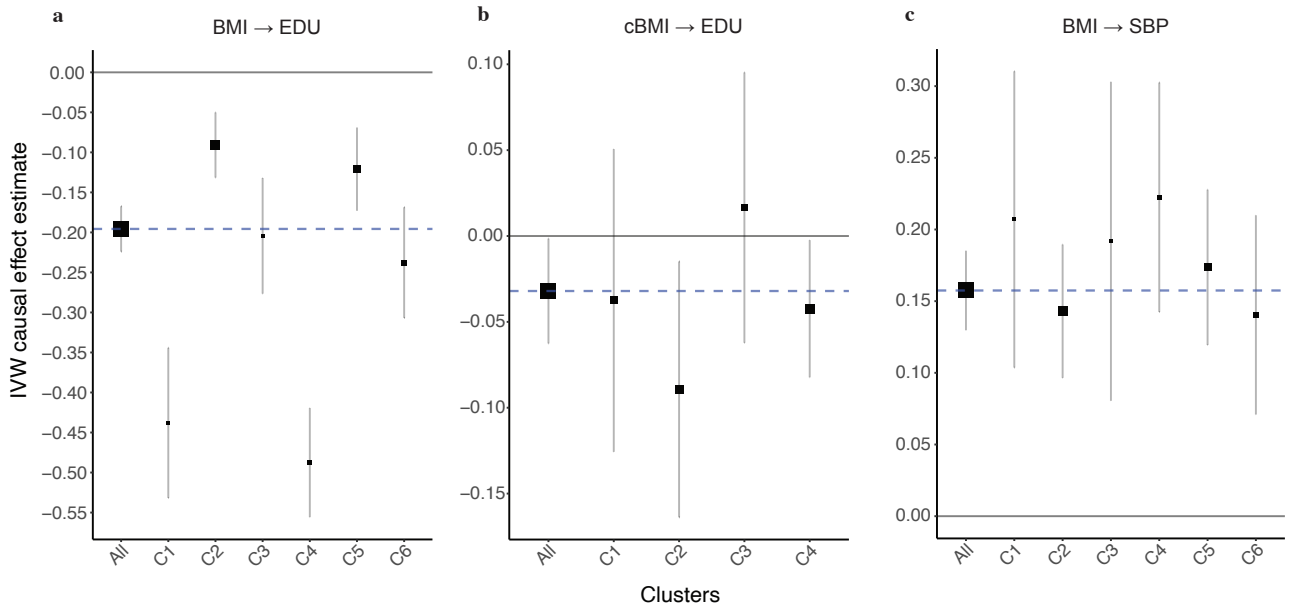

Figure 3: Forest plot of IVW causal effect estimate on outcome using either all exposure IVs or cluster-specific IVs. Panel **a** shows causal effect estimates of adult BMI on EDU, panel **b** proxy of childhood BMI (cBMI) on EDU, and panel **c** adult BMI on SBP. Horizontal error bars represent the 95% confidence interval. The blue vertical line represents the causal effect estimated using all BMI IVs. Box sizes of clusters represent the proportion of the number of IVs in each cluster to the total.

3.3 Post hoc analyses

To test the robustness of the PWC-MR results, we performed four additional analyses. First, we analysed the same exposure and outcome, but using sib-regression-based SNP effect sizes instead of SNP effects from GWAS of unrelated samples. Second, we replaced the exposure with childhood BMI and estimated its causal effect on EDU. Third, we replaced the outcome, EDU, with SBP. Finally, we executed a systematic search for confounders to include in a multivariable MR analysis.

3.3.1 Sib-regression MR

In Howe et al.(2022)[18], within-sibship (within-family) meta-analysed GWAS estimates were generated from 178,086 siblings across 19 cohorts. Using these effect estimates, MR was performed with BMI as exposure on multiple traits, including educational attainment. They used 418 independent and genome-wide significant genetic variants for BMI, and estimated its effect on EDU using IVW to be -0.05 (95% CI: $-0.09, -0.01$).

They also used jackknife to estimate the standard error of the difference between the sib-regression MR estimate and that of the GWAS of unrelated samples MR estimate, -0.19 (95% CI: $-0.22, -0.16$). Using the difference Z-score to generate a p-value for heterogeneity between the two estimates revealed a p-value < 0.001.

3.3.2 Causal effect of childhood BMI on Educational attainment

We used the UK Biobank trait 'Comparative body size at age 10' as a proxy for childhood BMI – a measure that has been validated against measured BMI in childhood[30,31] – for the exposure trait. Childhood BMI is expected to be less confounded by (parental) SEP compared to adult BMI and hence we expect to see a less biased causal effect on EDU. For this trait, we had 171 genome-wide significant SNPs that we used as IVs for the analysis. Of these, 16 SNPs were more strongly associated to traits other than childhood BMI and were thus excluded from further analysis. The standardised effect matrix of the remaining 155 SNPs across 461 traits was clustered into four clusters (yielding optimal AIC), each containing $37, 42, 32, 44$ IVs respectively (Supplementary Figure S4, Supplementary Table 6).
Analysing the trait enrichment for each cluster revealed only two clusters with high ER values: clusters #2 and #4 (Supplementary Figure S5, Supplementary Table 7). Cluster #2 had only two traits with ERs greater than 2, which were 'Number of fluid intelligence questions attempted within time limit' and 'Fluid intelligence score', whereas cluster #4 was highly enriched for body-measurement/fat-mass traits such as 'Waist circumference' and 'Whole body fat mass' (see Supplementary Figure S6). However, calculating the IVW causal effect estimate for each cluster and comparing it to the estimate calculated using all IVs revealed homogeneous causal effect estimates with a Q-statistic of 3.84 (p-value of 0.43) as seen in Figure 3**b** and Supplementary Table 8. Cluster #2 had a causal effect estimate of -0.09 (95% CI:-0.1638, -0.0148), and cluster #4 had a causal effect estimate of -0.04 (95% CI:-0.0823, -0.0024). Noteworthy is the finding that the IVs of cluster #2 were more heterogeneous than all the IVs combined. Thus, we obtained a massively attenuated causal effect of BMI on EDU, when childhood BMI is used as an exposure. Reassuringly, no SEP-enriched cluster emerged and the cluster specific causal effects were homogeneous.

3.3.3 Causal effect of BMI on SBP

To find further evidence that our approach does not always reveal distinct causal effects when the causal effect is non-null, we replaced EDU with SBP as outcome. Namely, we tested a well-established non-null causal relationship that is hypothesised to not be biased by pleiotropy or confounding: BMI's effect on SBP. Using the same six clusters previously obtained for BMI, we calculated the estimated causal effect of each of the clusters compared to using all the IVs combined on SBP. This revealed a homogeneous set of causal effect estimates (Q-test value of 4.49, p-value = 0.61), with the IVW estimate using all IVs being 0.15 (p-value = $1.09 \times 10^{-}28$) as seen in Figure 3**c** and Supplementary Table 5.

3.3.4 Systematic confounder search and MVMR analysis

Given our suspicion that the large BMI-EDU causal effect is driven by heritable confounders, we performed a systematic search to reveal traits that may be potential confounders. As described in the Methods section, the strength of the bidirectional effect of the traits on either the exposure or the outcome determined their categorisation. This led to the identification of 19 traits that were found to be candidate confounder traits (Supplementary Table 9). Matching the 19 confounder traits from this analysis to their respective ER across the six clusters from the previous analysis revealed higher ERs in cluster #1 and cluster #4 (see Supplementary Figure S7), which was associated with SEP-related traits. Noteworthy is that the traits labelled as candidate confounders were predominantly environmental exposures, such as 'Exposure to tobacco smoke outside home' and 'Transport type for commuting to job workplace: Cycle'.
Furthermore, these candidate confounder traits are attributed as *candidate* or *potential* confounders since they are most likely only genetic correlates of the true confounding traits of the

BMI-EDU relationship and not act as true confounders themselves.

To investigate the possible biasing effect that potential confounder traits can have on the causal relationship of BMI on EDU, we ran a stepwise MVMR on these 19 candidate confounder traits (Supplementary Table 9). During the creation of the Z-score matrix of SNPs and traits, only five traits had at least three genome-wide significant and independent SNPs whose effects could be used in the analysis. These five traits were: 'Time spent watching television (TV)', 'Usual walking pace', 'Past tobacco smoking', 'Frequency of tiredness / lethargy in last 2 weeks', and 'Average weekly beer plus cider intake'. Of these, only the first three had a significant causal effect estimate on EDU (p-value $< 0.05/5$) as estimated by stepwise MVMR, and were subsequently used as exposures alongside BMI in a standard MVMR analysis.

To ensure the strength of the IVs used in the MVMR analysis, we calculated the conditional F-statistic and the MVMR causal effect estimate of BMI given various combinations of the 3 remaining candidate confounder traits. We saw the expected trend of a decreasing conditional F-statistic with the addition of traits and their IVs to the analysis (see Supplementary Figure S8). We note that the causal effect estimate of BMI on EDU decreases when any combination of the candidate confounder traits is used with BMI as exposure in comparison to the univariable MR causal effect estimate of BMI on EDU (-0.19, p-value $= 7.11 \times 10^{-}41$). We settled on the combination of candidate confounder traits yielding a conditional F-statistic for BMI > 8, for which the corresponding causal effect estimates are reported in Table 1 below. This choice was a compromise between two sources of biases: weak instrument bias *vs* upward bias due to omitting relevant confounders.

Trait	Description	α estimate	SE	P-value	Conditional F-statistic
1070	Time spent watching television	-0.4077	0.0374	3.46E-23	10.80
1249	Past tobacco smoking	0.1394	0.0384	3.39E-04	15.74
21001	Body mass index (BMI)	-0.0633	0.0136	4.98E-06	56.89*

Table 1: MVMR analysis results of BMI and two candidate confounder traits on education. α: causal effect estimate. The conditional F-statistic column refers to BMI's calculated conditional F-statistic on each trait. The value in that column for BMI however, indicated by *, refers to BMI's F-statistic calculated when running a univariable MR with only BMI as exposure.

3.4 Relationship with other approaches

3.4.1 Comparison against MR-Clust

Other known IV clustering methods include MR-Clust[27], which attempts to cluster variants with similar causal effect estimates together following the hypothesis that exposures can affect an outcome by distinct causal mechanisms to varying extents. MR-Clust also accounts for the possibility of spurious clusters by assigning IVs with uncertain causal effect estimates to 'null' or 'junk' clusters.

We compared the k-means clustering of BMI IVs against that of MR-Clust with EDU as the outcome. The results revealed two main clusters as well as a 'null' cluster. Cluster #1 had 35 SNPs, 13 of which had an inclusion probability greater than 80%. Cluster #2 had 171 SNPs, 36 of which had an inclusion probability greater than 80%, and the remaining 142 SNPs were categorised into the 'null' cluster as seen in Supplementary Figure S9. The mean causal effect estimate of SNPs in cluster #1 was -0.496, whereas it was -0.246 for cluster #2. Searching for trait associations for the SNPs in each of the clusters revealed that body measurement traits like 'Arm fat mass' or 'Body fat percentage' are associated to SNPs in both clusters, while

SEP-related traits such as 'Fluid intelligence score' or 'Time spent watching television' were associated to more SNPs in cluster #1 than in cluster #2.

Comparing the SNP clustering between the k-means method against that of MR-Clust in Table 2 below, we see that cluster #1 in MR-Clust, which seems to be more strongly enriched for SEP traits than cluster #2, has SNPs that were similarly clustered in clusters #1 and #4 using k-means, matching their large negative causal effect of BMI on EDU. However, the same distinct comparison cannot be made for SNPs in cluster #2 of MR-Clust.
Of the 12 Fisher's exact tests performed to examine the contingency of SNPs in the two separate sets of clusters, four tests revealed a significant association: SNPs in cluster #1 of MR-Clust were significantly associated with SNPs in clusters #1, #2 (lean-mass traits), #4 (SEP-related traits) and #5 of the K-means clustering.

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
Cluster1	13	1	0	13	0	5
Cluster2	15	38	21	26	32	29
Null	4	59	14	2	37	15

Table 2: Cross table of BMI IVs clustered using K-means and MR-Clust.

3.4.2 Colocalisation analysis

With the aim of finding supporting evidence for the k-means clustering and enrichment analysis, we ran a genetic colocalisation analysis on BMI IVs and two types of tissue: subcutaneous adipose and brain, the results of which can be found in Supplementary Tables 10 and 11 respectively.

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
Adipose	9	9	14	3	6	5
Brain	3	3	4	1	2	4
Both	1	2	1	1	4	4
Neither	29	77	36	23	53	47

Table 3: Cross table indicating the number of genes whose expression colocalises in adipose/brain tissue with BMI. The colocalisation exercise was performed at loci-defined BMI IVs falling into particular clusters. Colocalisation was defined as the posterior probability of both GWAS and eQTL being associated is ≥ 0.8 in either brain or adipose tissue or both.

Running a set of Fisher's tests to compute the overlap between the membership of the SNPs in the six clusters and their tissue of colocalization did not reveal any association between clusters and tissues.

4 Discussion

We have developed a method that performs informative clustering of IVs by utilising their association with a large number of traits. Our use of PheWAS data to guide the clustering of IVs has revealed distinct mechanisms by which exposure effects could act on outcomes. For our exposure, BMI, six distinct clusters were obtained through optimal K-means clustering. These clusters had well-defined trait enrichments, with clusters matching SEP-related, substrate, and body measurement traits. Estimating individual causal effects of each cluster on EDU as an outcome revealed heterogeneous causal effect estimates which allowed us to further strengthen our suspicion that the MR estimate for the causal effect of BMI on EDU is upward biased when using population-based SNP effect size estimates due to confounding.

We note from MR analysis run using within-sibling GWAS data[18] that the causal effect estimate between BMI and EDU is -0.05 (95% CI: $-0.09, -0.01$), which is smaller than the causal effect estimate seen using population based GWAS data (-0.19, 95% CI: $-0.22, -0.16$). Investigating the various mechanisms or pathways through which BMI could have a causal effect estimate on EDU through trait-enrichment analysis has revealed notable causal effect estimates from two clusters: one with a strongly negative MR estimate whose trait enrichment reflects shared mechanisms with socio-economic factors, and another cluster with close to zero causal effect estimate enriched for lean-mass traits. MR has typically presented bias due to heterogeneous causal effects emerging via distinct pathways and bias due to confounding of the instrument-outcome association as being separate mechanisms. Here, we have illustrated that a pheWAS-based clustering approach can classify instruments into clusters, some of which correspond to different pathways, while others include IVs that are primarily confounder-associated. Our results have two major implications: 1) The lean-mass-related IV cluster indicated a close to zero causal effect of BMI on EDU. 2) We revealed that the SEP-related IVs suggest a sizeable negative effect of BMI on EDU.

In order to substantiate our findings, we performed several follow-up analyses. First, sib-regression based MR of BMI on EDU recapitulated the close-to-zero causal effect obtained for the body-mass specific cluster of IVs. This indicates that many IVs for adult BMI (from population-based GWAS) represent indirect (parental/dynastic) effects, which are associated rather with a rearing-related parental trait and not primarily with offspring BMI. Second, replacing adult BMI with childhood BMI (much less associated with SEP) as exposure in the PWC-MR analysis confirmed a negligible causal effect estimate (-0.03, p-value $= 0.04$), and the four emerging clusters showed homogeneous causal effect estimates indicating the lack of confounding or biasing effects. This comparison was supported by the growing evidence showing that genetic variants have varying effects on BMI or body size at different stages of life[32, 33], and that the UK Biobank proxy trait 'Comparative body size at age 10' captures childhood BMI well[30]. One of the four clusters was strongly enriched for body-measurement/fat-mass traits whereas the second most strongly enriched cluster had only two mildly enriched SEP-related traits. This finding means that as opposed to adult BMI, childhood BMI genetics are unrelated to childhood (i.e. parental) SEP. It is also interesting to note that although fat-mass traits are strongly enriched for using childhood BMI IVs alongside lean-mass traits, these same traits are less enriched for using adult BMI IVs. Thus, IVs associated with body-mass related traits seem to be underlying the true nominally significant and minuscule causal effect between BMI and EDU. Furthermore, out of the 41 adult BMI IVs that make up cluster #4 (SEP-related traits), only three were found to be in LD with childhood BMI IVs in cluster #2 (enriched for two SEP-related traits).

In Howe et al. (2022), assortative mating, dynastic effects and population stratification were all considered candidate mechanisms for biased population-based GWAS effect estimates. Given

our observations, a possible explanation is a dynastic effect of parental SEP traits acting as a confounder on both adult EDU and adult BMI (as seen in Figure 1**a**). This effect is direct on adult EDU but could affect adult BMI indirectly through either of two ways or both to a certain extent: (i) Parental SEP has a direct effect on the offspring's SEP as an adult, which in turn has an effect on offspring adult BMI[34], or (ii) parental SEP – as an indicator of childhood social circumstances – may have an effect through this on the offspring's (adult) BMI.

To explore the relevance of the obtained six clusters of IVs, we replaced EDU with SBP as the outcome of interest since within-sibling GWAS MR results showed no difference when compared to population GWAS MR results, indicating that there seems to be no bias in the causal effect estimate due to pleiotropy or confounding. Our analysis revealed that for the six clusters attributed to BMI, their causal effect estimate on SBP was homogeneous with the estimate using all SNPs (0.16, p-value = $1.09 \times 10^{-}28$). As there is no significant heterogeneous effects and all the cluster causal effects agree, we can conclude that there is no other confounding effects biasing the causal effect estimate. It is reassuring to note that our PWC-MR approach does not always seek to identify distinct causal effects, confirming that confounding mechanisms are specific to certain exposure-outcome pairs.

Finally, our systematic confounder search coupled with stepwise MVMR has pinpointed TV watching and smoking as two candidate confounder traits (maybe acting as a correlate of parental SEP) that may bias standard MR analysis of the BMI-EDU relationship: upon accounting for these two traits, BMI exhibits strongly attenuated causal effect on EDU.

Comparing our method to other IV clustering methods such as MR-Clust does not reveal strong concordance in the findings. MR-Clust takes as input the exposure and outcome effects as well as their standard errors and attempts to cluster the exposure IVs based on the possible similarity between each IV's causal effect. When using BMI and EDU as exposure and outcome respectively, MR-Clust revealed two main clusters alongside a null cluster. Both of the clusters were enriched for a variety of traits including body-measurement traits, both lean- and fat-mass, as well as SEP-related traits. The causal effect estimates of both clusters were strongly negative, similar to using all IVs in an MR analysis for this trait pair.
The most apparent difference between the clustering of our method and that of MR-Clust is our use of external information for the exposure and our clustering of IVs independently of the outcome or the individual MR causal effects of the IVs. By clustering based on the PheWAS data of the exposure IVs and various other traits, we can reveal possible pathways and mechanisms through which the exposure manifests, independently of any outcome.

Another comparable clustering method by Grant et al.[35] uses genetic variant associations with a set of traits to identify groups of IVs with similar biological mechanisms. Their method, NAvMix, uses a directional clustering algorithm and includes a noise-cluster to increase robustness to outliers. NAvMIX is demonstrated on BMI IVs and their associations to nine lifestyle or cardio-metabolic traits that have been previously shown to be related to BMI. Their results revealed 5 distinct clusters where they were able to identify a metabolically healthy obesity cluster that also had a small MR causal effect on coronary heart disease (CHD). However, we were unable to run their method using our data due to convergence issues arising when the number of traits used for PheWAS association increases. This comparison also highlights that the traits we include in the pheWAS analysis (and the subsequent clustering) have an important role in which biological mechanisms we can detect. For example, our analysis did not pick up the metabolically healthy obesity cluster, potentially because waist-to-hip ratio and other subcutaneous-vs-visceral fat proxy-traits were not included among the 408 selected phenotypes due to our filtering on genetic correlation with BMI ($r^g < 0.75$). Without such filtering, PWC-MR reveals 5 clusters with significantly heterogeneous causal effects on EDU. These five clusters

are very similar to the original six, with the original cluster #1 getting diffused into the other clusters. Reassuringly, the cluster that is strongly enriched for SEP-related traits has a large negative causal effect estimate of -0.53 (95% CI: $-0.59, -0.48$), whereas the cluster that is most enriched for body-measurement/fat-mass traits still had an attenuated causal effect of -0.10 (95% CI: $-0.14, -0.06$).

Furthermore, we attempted to consolidate our findings of the k-means clustering and enrichment analysis by running a genetic colocalisation analysis on the 324 clustered BMI IVs and both subcutaneous adipose and brain tissue. Unfortunately, we do no find an association between the cluster memberships of the IVs and their signal colocalization in brain or adipose tissue, possibly due to the limited evidence of colocalization in either tissue for most loci.

Our method has its own set of limitations: first, we are limited by the availability of traits with PheWAS data to support our informative clustering of IVs. This may lead to a failure in identifying key pathways and thus missing clusters representing important subgroup (mediator/subphenotype/confounder). Second, although it is not the most ideal handling of data, our binary traits are treated as continuous ones in our analysis. In large samples, linear and logistic regression effect estimates correlate very strongly and hence, it is likely that this choice did not impact the clustering[36]. Third, although we have attempted to minimise the arbitrary choice of parameters in our analysis, the genetic correlation threshold that determines which traits are too similar to the exposure and outcome trait is arbitrarily set at 0.75 for BMI and EDU and could be modified, but the emerging clusters may change as a consequence. Similarly, some p-value thresholds and type I error rate control was set at 5%, which may be viewed as arbitrary. Fourth, the identified potential confounder traits used in the MVMR analysis act as simple proxies for true confounders. For example, exposure to current tobacco smoking or TV watching can be highly (genetically) correlated to the same or a similar exposure during early life (or even proxy a parental trait), hence it is rather the earlier version of the exposure which is likely to be the true confounder. Our proxy confounders were simply nuisance variables, their only role was to see the remaining causal effect of BMI on EDU upon conditioning on them. Lastly, we acknowledge that there are several other tests[37] that could be used in place of a t-test when excluding SNPs more strongly associated to other traits than our exposure or different MR methods used in our systematic confounder search, however both of these were simple exclusion or pre-selection steps that have very little impact on the outcome of the results.

Acknowledgements

We are grateful for the useful discussions on MVMR with Eleanor Sanderson. This research has been conducted using the UK Biobank Resource under Application Number 16389. Z.K. was funded by the Swiss National Science Foundation (310030_189147 and 32003B_173092). GH and GDS work within the MRC Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/1). Payments were made to the institution. For computations, we used the CHUV HPC cluster.

Author contributions

L.D. and Z.K. conceived and designed the project. Z.K. supervised all statistical analyses. L.D. implemented the research and performed the analyses. L.D. and Z.K. prepared the first draft of the manuscript. L.D., Z.K., G.H. and G.D.S. contributed to the review and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Code availability

The source code for this work can be found on `https://github.com/LizaDarrous/PheWAS-cluster`.

References

[1] Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers *1*, 59.

[2] Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., et al. (2022). A saturated map of common genetic variants associated with human height. Nature *610*, 704–712.

[3] Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet *50*, 1112–1121.

[4] Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., et al. (2022). Mendelian randomization. Nature Reviews Methods Primers *2*, 6.

[5] Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics *23*, R89–R98.

[6] Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol *37*, 658–665.

[7] Watanabe, K., Stringer, S., Frei, O., UmićevićMirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics *51*, 1339–1348.

[8] Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. Int J Epidemiol *44*, 512–525.

[9] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat Genet *52*, 740–747.

[10] Darrous, L., Mounier, N., and Kutalik, Z. (2021). Simultaneous estimation of bi-directional causal effects and heritable confounding from gwas summary statistics. Nature Communications *12*, 7274.

[11] Young, A. I., Benonisdottir, S., Przeworski, M., and Kong, A. (2019). Deconstructing the sources of genotype-phenotype associations in humans. Science *365*, 1396–1400.

[12] Robinson, M. R., Kleinman, A., Graff, M., Vinkhuyzen, A. A. E., Couper, D., Miller, M. B., Peyrot, W. J., Abdellaoui, A., Zietsch, B. P., Nolte, I. M., et al. (2017). Genetic evidence of assortative mating in humans. Nature Human Behaviour *1*, 0016.

[13] Howe, L. J., Lawson, D. J., Davies, N. M., St. Pourcain, B., Lewis, S. J., Davey Smith, G., and Hemani, G. (2019). Genetic evidence for assortative mating on alcohol consumption in the uk biobank. Nature Communications *10*, 5039.

[14] Haworth, S., Mitchell, R., Corbin, L., Wade, K. H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Davey Smith, G., et al. (2019). Apparent latent

structure within the uk biobank sample has implications for epidemiological analysis. Nat Commun *10*, 333.

[15] Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., and Davey Smith, G. (2019). Within family mendelian randomization studies. Hum Mol Genet *28*, R170–R179.

[16] Benyamin, B., Visscher, P. M., and McRae, A. F. (2009). Family-based genome-wide association studies. Pharmacogenomics *10*, 181–190.

[17] Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., et al. (2020). Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. Nature Communications *11*, 3519.

[18] Howe, L. J., Nivard, M. G., Morris, T. T., Hansen, A. F., Rasheed, H., Cho, Y., Chittoor, G., Ahlskog, R., Lind, P. A., Palviainen, T., et al. (2022). Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. Nature Genetics *54*, 581–592.

[19] Neale Lab (2018). UK BioBank - round 2. `http://www.nealelab.is/uk-biobank/`.

[20] Hemani, G., Zheng, J., Elsworth, B., Wade, K., Baird, D., Haberland, V., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The mr-base platform supports systematic causal inference across the human phenome. eLife *7*, e34408.

[21] Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using gwas summary data. PLOS Genetics *13*, 1–22.

[22] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) *28*, 100–108.

[23] Mounier, N. and Kutalik, Z. (2022). Bias correction for inverse variance weighting mendelian randomization. bioRxiv.

[24] Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. Stat Med *36*, 1783–1802.

[25] Mounier, N. and Kutalik, Z. (2020). bGWAS: an R package to perform Bayesian genome wide association studies. Bioinformatics *36*, 4374–4376.

[26] Sanderson, E., Spiller, W., and Bowden, J. (2021). Testing and correcting for weak and pleiotropic instruments in two-sample multivariable mendelian randomization. Statistics in Medicine *40*, 5434–5452.

[27] Foley, C. N., Mason, A. M., Kirk, P. D. W., and Burgess, S. (2020). MR-Clust: clustering of genetic variants in Mendelian randomization with similar causal estimates. Bioinformatics *37*, 531–541.

[28] Leyden, G. M., Shapland, C. Y., Davey Smith, G., Sanderson, E., Greenwood, M. P., Murphy, D., and Richardson, T. G. (2022). Harnessing tissue-specific genetic variation to dissect putative causal pathways between body mass index and cardiometabolic phenotypes. The American Journal of Human Genetics *109*, 240–252.

[29] Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet *10*, e1004383.

[30] Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K., and Davey Smith, G. (2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. BMJ *369*.

[31] Brandkvist, M., Bjørngaard, J. H., Ødegård, R. A., Åsvold, B. O., Davey Smith, G., Brumpton, B., Hveem, K., Richardson, T. G., and Vie, G. Å. (2020). Separating the genetics of childhood and adult obesity: a validation study of genetic scores for body mass index in adolescence and adulthood in the HUNT Study. Human Molecular Genetics *29*, 3966–3973.

[32] Alves, A. C., Silva, N. M. G. D., Karhunen, V., Sovio, U., Das, S., Taal, H. R., Warrington, N. M., Lewin, A. M., Kaakinen, M., Cousminer, D. L., et al. (2019). Gwas on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult bmi. Science Advances *5*, eaaw3095.

[33] Richardson, T. G., Power, G. M., and Davey Smith, G. (2022). Adiposity may confound the association between vitamin d and disease risk - a lifecourse mendelian randomization study. Elife *11*.

[34] Blane, D., Hart, C. L., Davey Smith, G., Gillis, C. R., Hole, D. J., and Hawthorne, V. M. (1996). Association of cardiovascular disease risk factors with socioeconomic position during childhood and during adulthood. BMJ *313*, 1434–1438.

[35] Grant, A. J., Gill, D., Kirk, P. D. W., and Burgess, S. (2022). Noise-augmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity. PLOS Genetics *18*, 1–24.

[36] Pedersen, E. M., Agerbo, E., Plana-Ripoll, O., Steinbach, J., Krebs, M. D., Hougaard, D. M., Werge, T., Nordentoft, M., Børglum, A. D., Musliner, K. L., et al. (2022). Adult: An efficient and robust time-to-event gwas. medRxiv.

[37] Brown, B. C. and Knowles, D. A. (2021). Welch-weighted egger regression reduces false positives due to correlated pleiotropy in mendelian randomization. Am J Hum Genet *108*, 2319–2335.

Supplementary Information

Supplementary Figures

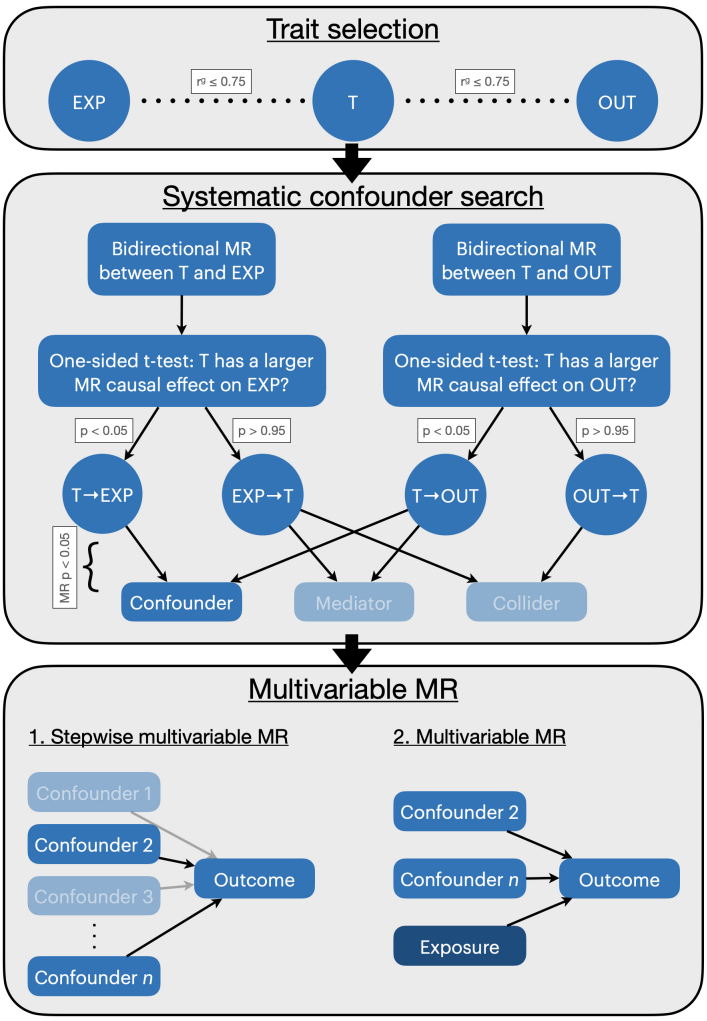

Figure S1: Flow diagram representing a complimentary approach to PWC-MR where a systematic candidate confounder trait search is performed. These canbdidate confounder traits are defined as having an effect on both the exposure and the outcome. In the third step, a stepwise multivariable MR of the candidate confounder traits is performed to select those with a strong effect on the outcome. They are then added with the original exposure to a standard MVMR and the multivariable causal effect on the outcome is estimated .Acronyms: EXP - exposure, OUT - outcome, T - trait, p: t-test p-value; MR P: MR p-value

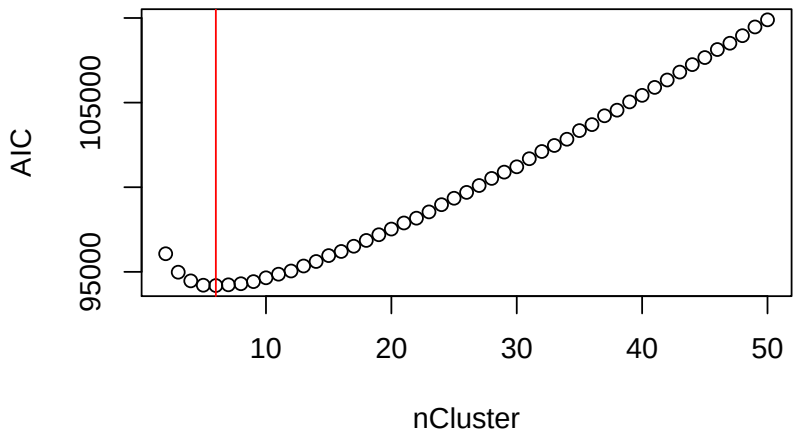

Figure S2: Dot plot representing the corresponding Akaike Information Criterion scores across varying K-means centres for BMI. K-means centres vary from 2 to 50 clusters. The red vertical line represents the number of centres/cluster with the lowest score.

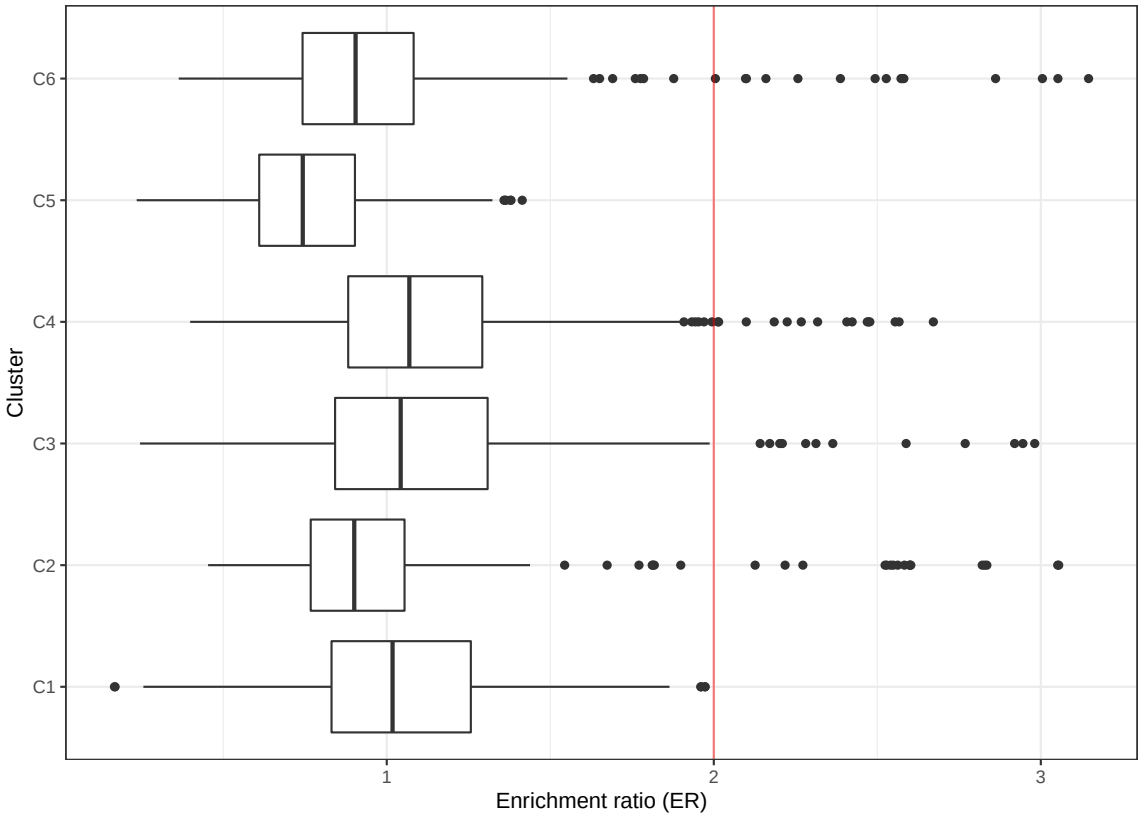

Figure S3: Boxplot showing the enrichment ratio of all traits in each cluster. BMI IVs have been clustered into 6 clusters using K-means. The enrichment ratio of each trait calculated using the cluster-specific IVs is shown in the barplot. In the boxplots, the lower and upper hinges correspond to the first and third quartiles, the middle bar corresponds to the median, whereas the upper whisker is the largest data point smaller than 1.5∗ inter-quartile range above the third quartile. The lower whisker is defined analogously.

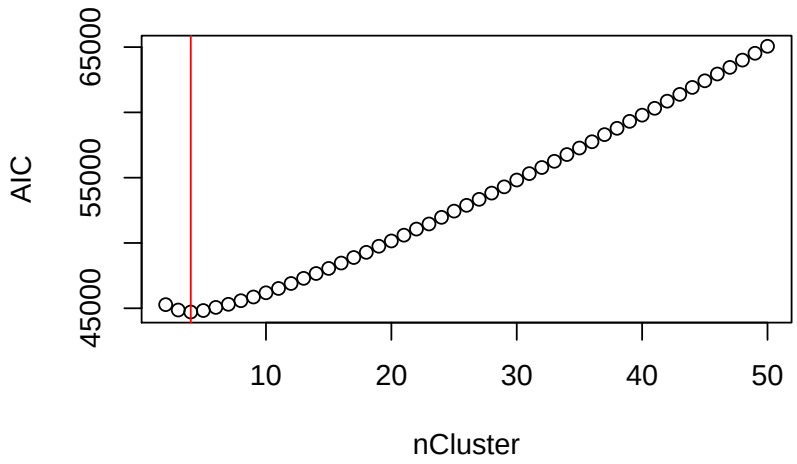

Figure S4: Dot plot representing the corresponding Akaike Information Criterion scores across varying K-means centres for child BMI. K-means centres vary from 2 to 50 clusters. The red vertical line represents the number of centres/cluster with the lowest score.

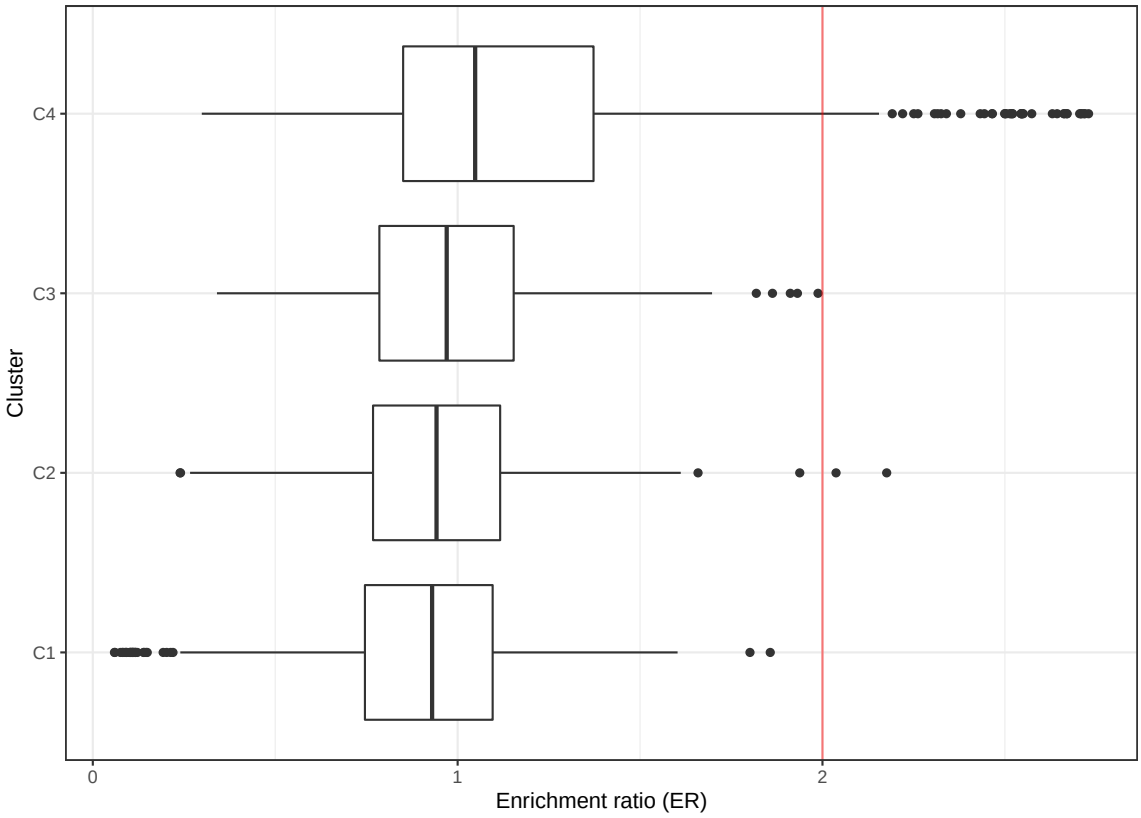

Figure S5: Boxplot showing the enrichment ratio of all traits in each cluster. Child BMI IVs have been clustered into 4 clusters using K-means. The enrichment ratio of each trait calculated using the cluster-specific IVs is shown in the barplot. In the boxplots, the lower and upper hinges correspond to the first and third quartiles, the middle bar corresponds to the median, whereas the upper whisker is the largest data point smaller than 1.5∗ inter-quartile range above the third quartile. The lower whisker is defined analogously.

Figure S6: Heatmap of the enrichment ratio of the top 10 traits in each cluster. Body size at age 10 is used as a proxy exposure trait for child BMI. K-means clustering revealed 4 clusters with the following trait enrichment ratios.

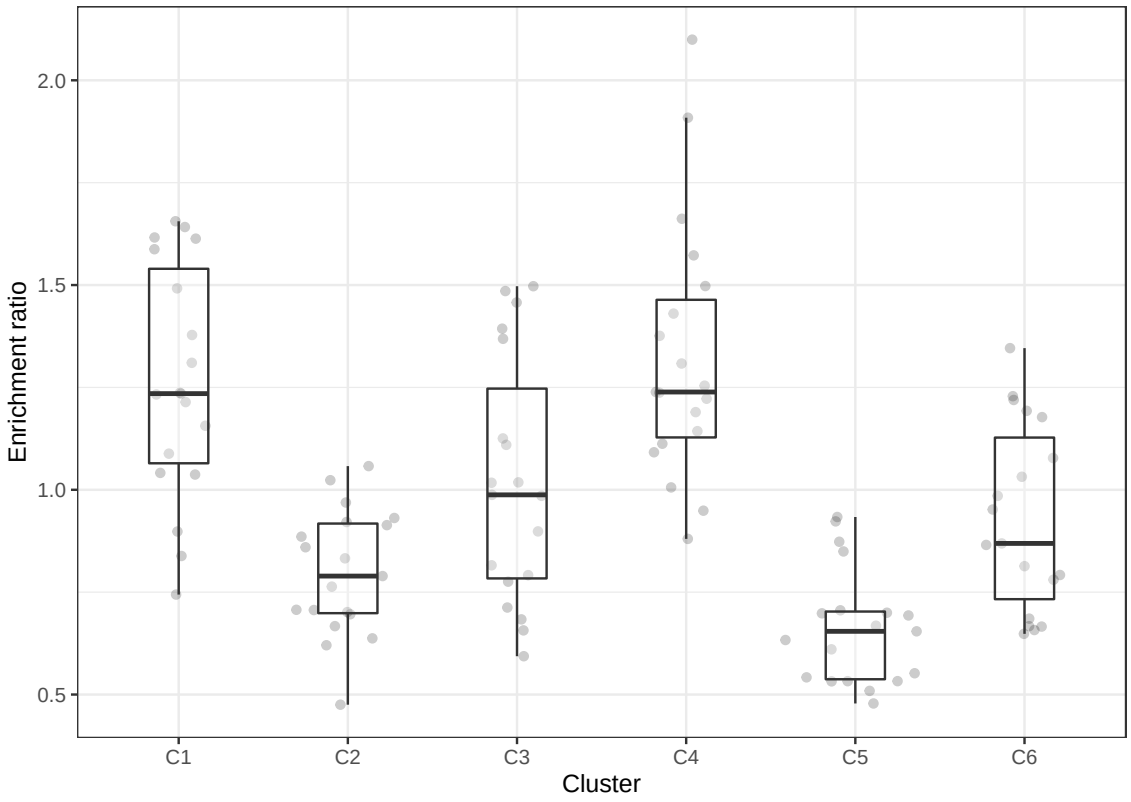

Figure S7: Boxplot showing the ER for confounder traits across the clusters. Confounder traits were categorised in a systematic search. In the boxplots, the lower and upper hinges correspond to the first and third quartiles, the middle bar corresponds to the median, whereas the upper whisker is the largest data point smaller than 1.5∗ inter-quartile range above the third quartile. The lower whisker is defined analogously.

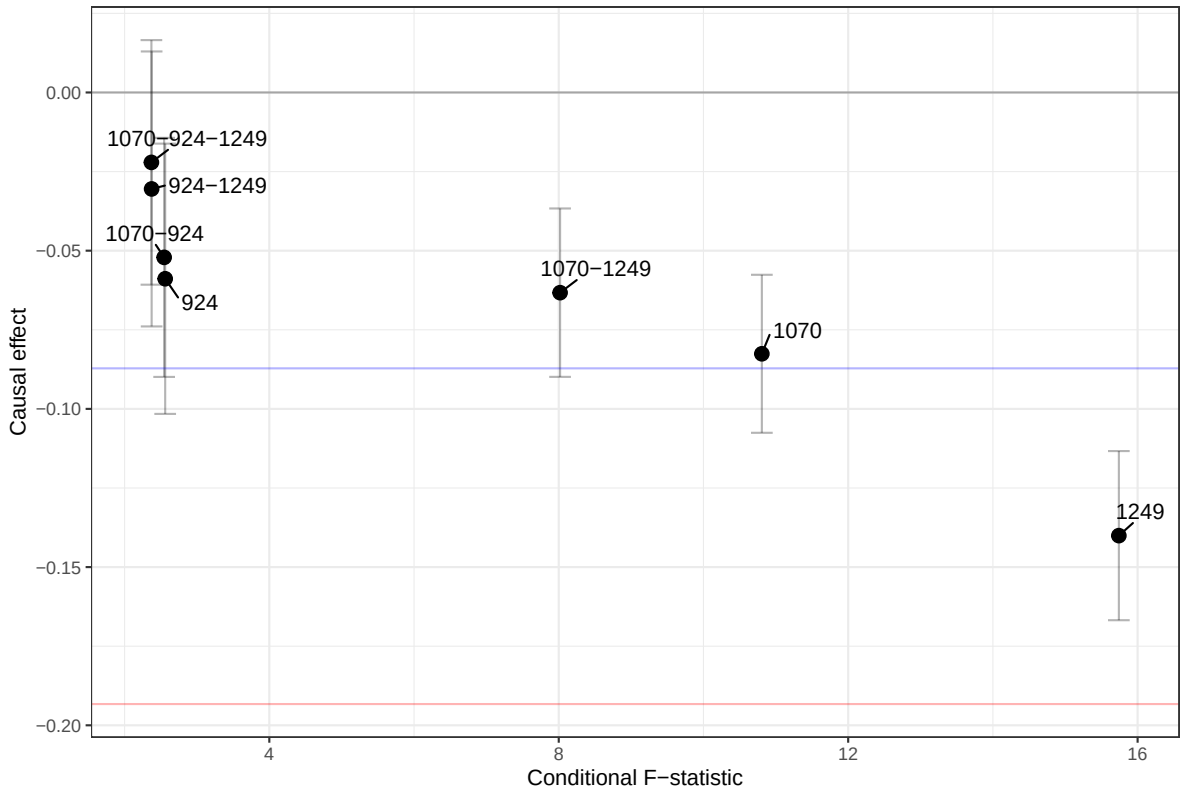

Figure S8: Dot plot showing the causal effect estimate of BMI on EDU conditional on various combinations of three candidate confounder traits. The error bars represent the 95% CI. The blue horizontal line represents the observational correlation between BMI and EDU, whereas the red horizontal line represents the univariate causal effect estimate of BMI on EDU. Trait 1070: 'Time spent watching television (TV)', trait 924: 'Usual walking pace', trait 1249: 'Past tobacco smoking'.

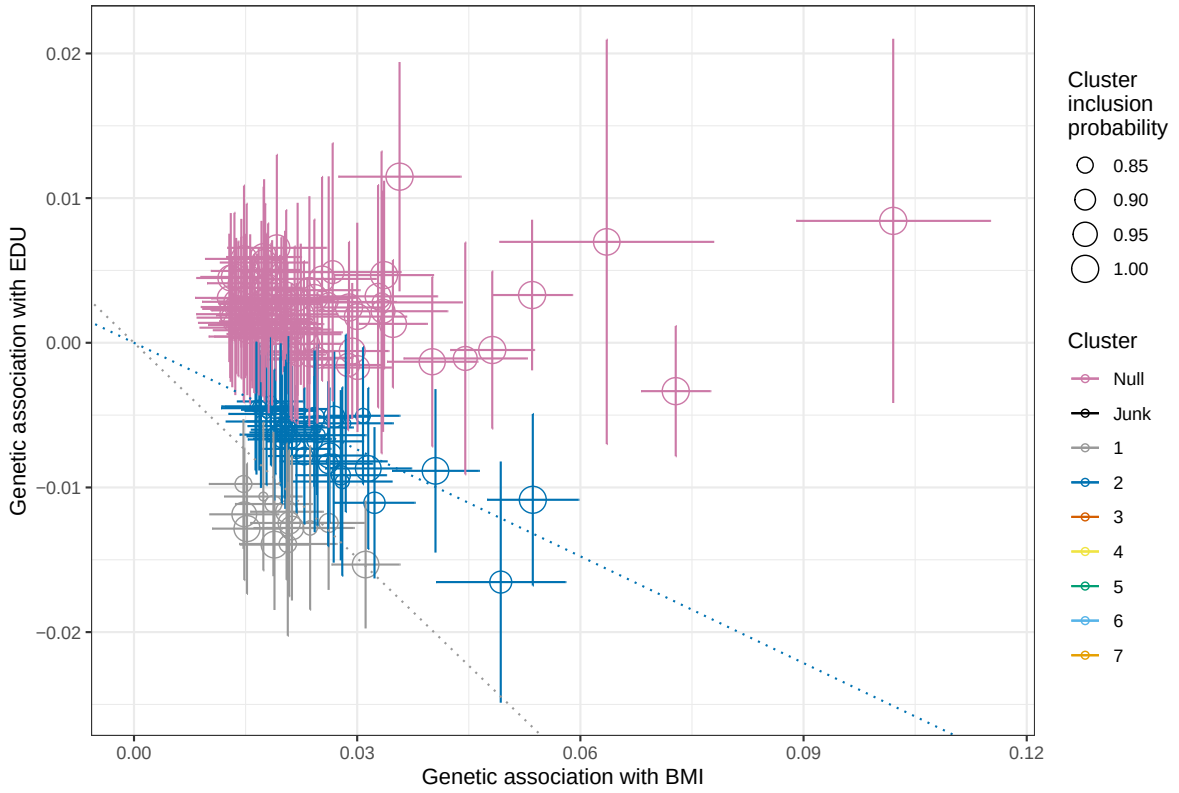

Figure S9: Dot plot showing the genetic association of IVs with the exposure: BMI, and the outcome: EDU. The exposure IVs have been clustered using MR-Clust based on their similarity in causal effect estimates. MR-Clust has revealed 2 main clusters for BMI's causal effect on EDU as well as a 'null' cluster. The IVs plotted have a cluster inclusion probability greater than or equal to 80%. The slopes represent the causal effect estimate of each cluster.