MR-AHC: A fast, efficient and robust method for two-sample summary data Mendelian randomization based on agglomerative hierarchical clustering

Xiaoran Liang[a] ,*Ninon Mounier[a], Nicolas Apfel[b], Sara Khalid[c], Timothy M Frayling[a], and Jack Bowden[a,d] on behalf of the GEMINI Consortium

[a]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK
[b]Department of Economics, University of Southampton, Southampton, UK
[c]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[d]Novo Nordisk Genetics Centre of Excellence, Oxford, UK

March 18, 2023

Abstract

Mendelian randomization (MR) is an epidemiological approach that uses genetic variants as instrumental variables for estimating the causal effect of a modifiable but likely confounded exposure on an outcome. Standard MR usually assumes that all included genetic variants are valid instruments and there is a single homogeneous causal effect of the exposure on the outcome. We allow violations of both assumptions such that the variants can be divided into clusters identifying distinct causal effects driven by different biological mechanisms and/or horizontal pleiotropy. We adapted the Agglomerative Hierarchical Clustering (AHC) method developed for individual-level data to the summary data MR setting, enabling the detection of such variant clusters using only genome-wide summary statistics. We also extend the method to handle two outcomes and a common exposure to aid investigation of the mechanisms of multimorbidity. We conduct Monte Carlo simulations to evaluate the performance of our 'MR-AHC' algorithm compared to the existing MR-Clust method, showing that it is both reliable and computationally efficient: it detects variant clusters with high accuracy and is much faster than MR-Clust. In an applied example, we use our method to analyze the causal effects of high body fat percentage on a pair of well-known multimorbid conditions, type 2 diabetes and osteoarthritis, discovering distinct variant clusters reflecting the heterogeneous shared pathways.

Keywords: Mendelian randomization; clustering analysis; robust MR; hierarchical clustering; heterogeneous causal effects; multimorbidity.

*Correspondence: x.liang2@exeter.ac.uk

1 Introduction

Mendelian randomization (MR) is a widely used method in epidemiology that uses genetic variants (usually in the form of single nucleotide polymorphisms, SNPs) as instrumental variables (IV) for estimating the causal effect of a potentially confounded exposure on an outcome [1]. A genetic variant is required to satisfy three IV conditions to be a valid instrument for testing causality: it is robustly associated with the exposure (IV1); it is independent of unobserved confounders of the exposure-outcome relationship (IV2); and it must only affect the outcome through the exposure (IV3). IV2 and IV3 can jointly be referred to as the exclusion restriction. In addition to instrument validity, a further assumption is required to estimate the causal effect, namely causal homogeneity. For example, in the case of a continuous exposure we could assume that the effect of a unit increase in the exposure is the same for all individuals and this same, single causal effect can be consistently estimated by each SNP. Figure 1 illustrates the standard setup assumed in an MR study, here G_j, X, Y and U represent a (single) genetic instrument, exposure, outcome and unmeasured confounders respectively. Solid lines and dashed lines indicate the relationships assumed to be present and absent by assumptions IV1-IV3, and by the causal homogeneity assumption there is a single causal pathway from the exposure to the outcome that is not affected by G_j.

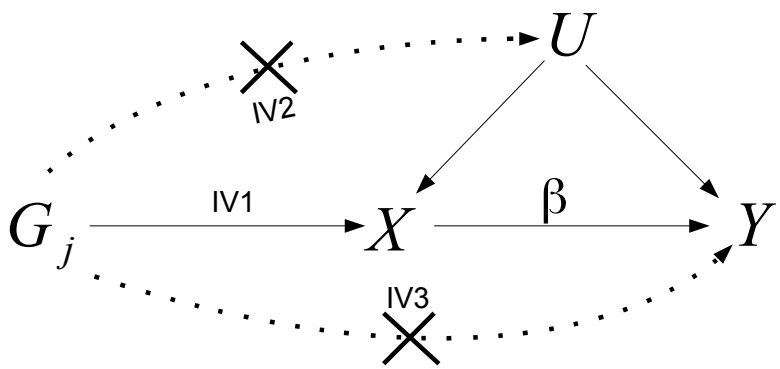


Figure 1: Causal diagram representing the assumed relationship between a genetic variant G_j, and exposure X and outcome Y within a Mendelian randomization study. β represents the causal effect of X on Y.

MR studies have traditionally been conducted using individual-level data on which the requisite genetic, exposure and outcome information is available. However, in recent years it has become more common to base the analysis on summary statistics of genetic associations gleaned from one or more genome-wide association studies (GWAS) [2]. With such summary statistics, SNP-specific causal estimates are first obtained as the ratio of their SNP-outcome and SNP-exposure associations, which we refer to as the Wald ratio estimate, and then the full set of estimates are combined using the principles of meta-analysis, the most common approach being inverse-variance weighted (IVW) meta-analysis [3]. This approach can be viewed conceptually

as finding the single best fitting line through the origin in a regression of the SNP-outcome association estimates on the SNP-exposure association estimates. If all the SNPs are valid instruments and the causal homogeneity assumption is satisfied, all the SNPs will target the same true causal effect and the IVW estimate will be unbiased and statistically efficient [4]. However, if one or more of the IV assumptions are violated, the basic approach may yield biased results.

For example, a major concern in MR is violation of the exclusion restriction due to *pleiotropy*, the general phenomenon whereby a SNP affects multiple traits [5] and therefore is associated with the outcome through pathways other than the exposure. As noted by Shapland et al. [6], if across all SNPs these pleiotropic effects are correlated with the strength of their SNP-exposure associations (e.g. because they influence a genetic confounder between the exposure and outcome), this correlated pleiotropy would be a violation of the InSIDE (Instrument Strength Independent of Direct Effect) assumption [4], and can give rise to multiple slopes in the summary data, even if the exposure has a single homogeneous effect on the outcome. This is illustrated by the causal diagram in Figure 2 (i) and the summary data scatter plot in Figure 2 (iii). The black slope represents the true causal effect identified by a group of non-pleiotropic SNPs G_2 and the red slope represents a biased effect identified by a group of pleiotropic SNPs G_1. The blue line corresponds to the average slope identified by the IVW estimate when both groups of SNPs (G_1 and G_2) are used, which is clearly not satisfactory.

Figure 2 (ii) shows a separate causal diagram representing violation of the causal homogeneity assumption, illustrating the scenario where the exposure X has been incorrectly defined as a single quantity when it is in fact made up of two or more sub-components (here X_1 and X_2) with different causal effects [7, 8]. For example, it is suspected that general adiposity exerts a heterogeneous causal effect on type 2 diabetes (T2D) depending on its location in the body (e.g. if it is around organs or in the limbs) [9]. This very different phenomenon could also give rise to the same summary data scatter plot in Figure 2 (iii). SNP group G_1 and G_2 could now represent instruments for sub-exposures X_1 and X_2, each responsible for true but distinct causal effects represented by the black and red slopes respectively. We follow Iong et al. [8] by referring to both correlated pleiotropy and causal effect heterogeneity under the umbrella term of 'mechanistic heterogeneity', since both features can give rise to the appearance of multiple slopes in the summary data, and it is impossible to discern the root cause without further data, domain knowledge and modelling assumptions.

In the presence of mechanistic heterogeneity, an important observation is that the causal estimates obtained with all SNPs will be heterogeneous, and only SNPs belonging to the same causal pathway have similar, or clustered, causal estimates. To discover such variant clusters that drive distinct causal mechanisms, we extend the Agglomerative Hierarchical Clustering (AHC) method in Apfel and Liang [10], as proposed in the field of econometrics to detect

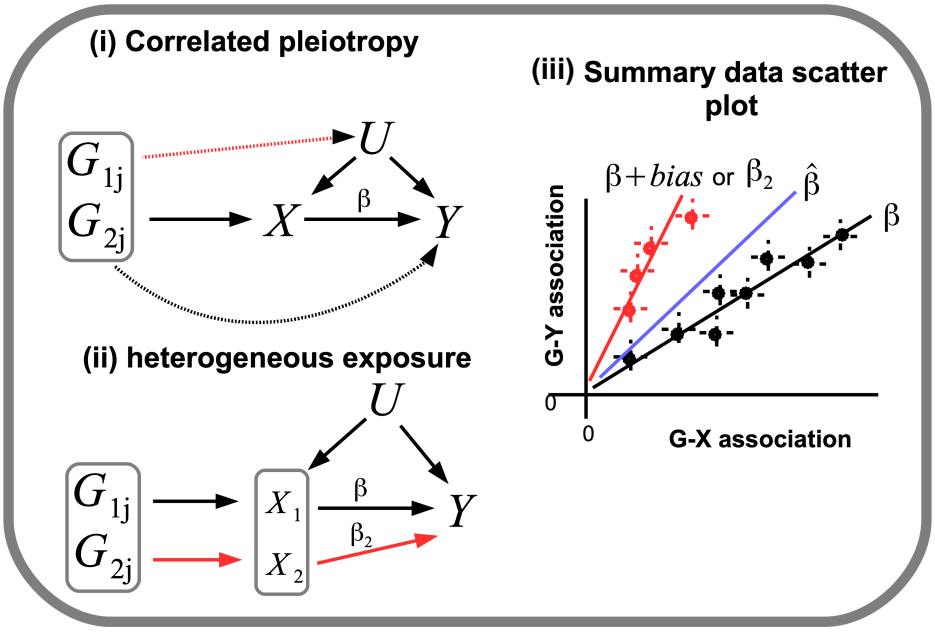


Figure 2: Two causes of SNP clustering in the summary data scatter plot: (i) correlated pleiotropy in a subset of SNPs and (ii) causal effect heterogeneity in the exposure.

instrument clusters, from the individual-level data setting to the two-sample summary data MR setting, which we call 'MR-AHC'. Specifically, we propose an adjusted version of Ward's minimum variance criterion for choosing the pair of clusters to merge at each step to account for the uncertainty of the clustering objects (i.e. the SNP-specific Wald estimates) due to being a function of summary data estimates rather than individual data points. This modification improves the statistical efficiency of the method considerably by involving the inverse-variance weight. We also propose a two-step extension of the method to account for outliers in the Wald estimates, which further improves the clustering accuracy. In addition to achieving accurate clustering results, our method is also computationally efficient and very fast to run. We provide easy-to-use R package that implements our method.

There are several existing methods for detecting the variant clusters in MR, for example, the MR-Clust method proposed by Foley et al. [7] and the MR-PATH method by Iong et al. [8]. Like MR-AHC, these two methods do not fix the number of clusters *a priori*, allowing the data to determine this important parameter. But notably, these methods only focus on the MR setting with one exposure and one outcome. We extend the MR-AHC method to the setting with two outcomes and a common exposure. This setting is of particular interest when studying the mechanisms of multimorbidity, the co-existence of two or more long-term conditions (LTCs) in an individual not by chance [11]. Existing studies suggest that many multimorbid LTCs have shared genetic risks [12], indicating an important role of genetic overlap between co-occurring LTCs when investigating the underlying mechanisms. The extended MR-AHC algorithm can be applied to investigate the shared but possibly heterogeneous causal pathways between a pair of

genetically correlated LTCs, whereby their genetic correlation can be attributed to the genetics associated with a common exposure that affects the LTC pair simultaneously.

The paper is structured as follows. In Section 2 we introduce the model setup for mechanistic heterogeneity in the context of MR. In Section 3 we introduce the MR-AHC algorithm and the extension of the method to handle two outcomes and a common exposure. In Section 4.1, we conduct Monte Carlo simulations to evaluate the performance of MR-AHC and compare with the MR-Clust method. In Section 4.2, we apply MR-AHC to explore the causal relationship between body fat percentage (BFP) and a pair of multimorbid conditions, T2D and osteoarthritis (OA). Section 5 provides a discussion and points to further work.

2 Materials and methods

2.1 Model setup

In order to motivate our method, and to provide a basis for subsequent simulation studies testing its performance, we assume a general linear IV model which takes pleiotropy and multiple exposure sub-components into account. Let the exposure be denoted by $X = X_1 + ... + X_k$ where X_k is the k-th sub-component in X and $k = 1, ..., K$. Let the observed outcome be denoted by the scalar Y. The vector $\mathbf{G} = \left(G_1, ..., G_p\right)'$ is used to denote the p genetic variants used as instruments. Throughout the paper we use 'SNPs' to represent genetic instruments. We then have the following linear structural model:

$$U = \sum_{j=1}^{p} \eta_j G_j + \epsilon_U, \tag{1}$$

$$X_k = \sum_{j=1}^{p} \delta_{kj} G_j + q_{xk} U + \epsilon_{Xk}, \tag{2}$$

$$X = \sum_{k=1}^{K} X_k, \tag{3}$$

$$Y = \sum_{k=1}^{K} \beta_k X_k + \sum_{j=1}^{p} \psi_j G_j + q_y U + \epsilon_Y, \tag{4}$$

where

- U represents the unobserved confounding between Y and any sub-component of X, with the strength of confounding determined by parameters q_{xk} and q_y;
- ϵ_U, ϵ_{Xk} and ϵ_Y are independent error terms affecting U, X_k and Y respectively;

- $\beta_1, ..., \beta_K$ represent the potentially heterogeneous causal effects of X on Y.

It follows from (4) that the pleiotropic effect of G_j on Y is $\alpha_j = \psi_j + q_y \eta_j$. Assuming $q_y \neq 0$ and $q_{xk} \neq 0$, then InSIDE violating SNPs are those with $\eta_j \neq 0$. Combining (1) and (2), we obtain the reduced-form relationship between X_k and G_j as $X_k = \sum_{j=1}^{p} \gamma_{kj} G_j + \xi_{Xk}$, where the total effect of G_j on X_k is $\gamma_{kj} = \delta_{kj} + q_{xk}\eta_j$. It follows from (3) that the overall effect of G_j on exposure X is

$$\gamma_j = \sum_{k=1}^{K} \gamma_{kj} = \sum_{k=1}^{K} \delta_{kj} + \eta_j \sum_{k=1}^{K} q_{xk}. \tag{5}$$

We assume that the relevance condition IV1 is satisfied at the scale of the overall exposure X so that $\gamma_j \neq 0$ for $j = 1, ..., p$. Similarly, we can obtain the reduced-form between Y and G_j as $Y = \sum_{j=1}^{p} \Gamma_j G_j + \xi_Y$, where

$$\Gamma_j = \sum_{k=1}^{K} \beta_k \gamma_{kj} + \alpha_j = \sum_{k=1}^{K} \beta_k \delta_{kj} + \eta_j \sum_{k=1}^{K} \beta_k q_{xk} + \psi_j + q_y \eta_j, \tag{6}$$

by plugging (1) and (2) into (4). Then the SNP-specific causal estimand, and subsequent estimate is:

$$\beta_j = \frac{\Gamma_j}{\gamma_j} = \frac{\sum_{k=1}^{K} \beta_k \gamma_{kj} + \alpha_j}{\sum_{k=1}^{K} \gamma_{kj}} \quad \text{and } \widehat{\beta}_j = \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}.$$

where $\widehat{\Gamma}_j$ and $\widehat{\gamma}_j$ are estimates for Γ_j and γ_j. As the sample size n goes to infinity, each $\widehat{\beta}_j$ converges to their Wald ratio estimand β_j, which is the causal effect identified by SNP G_j. In the simple case where G_j only instruments sub-component X_k, the SNP-specific causal estimand then becomes:

$$\beta_j = \frac{\beta_k \gamma_{kj} + \alpha_j}{\gamma_{kj}} = \beta_k + \frac{\alpha_j}{\gamma_{kj}}. \tag{7}$$

2.2 Estimating the causal effect in two-sample summary data MR

We will assume that the following summary statistics $\left\{\widehat{\Gamma}_j, se\left(\widehat{\Gamma}_j\right)\right\}_{j=1}^{p}$ and $\left\{\widehat{\gamma}_j, se\left(\widehat{\gamma}_j\right)\right\}_{j=1}^{p}$ are available from two independent cohorts who are nevertheless part of the same underlying population, in the sense that their exposure and outcome data are both generated from Models (1)-(4). We will also assume that the instruments are themselves mutually uncorrelated (i.e. not in linkage disequilibrium). From this we make the following normality assumption:

Assumption 1.

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\Gamma}_j \\ \widehat{\gamma}_j \end{pmatrix} - \begin{pmatrix} \Gamma_j \\ \gamma_j \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{Yj}^2 & 0 \\ 0 & \sigma_{Xj}^2 \end{pmatrix}\right)$$

for $j = 1, ..., p$. It follows that $\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right) \xrightarrow{d} N\left(0, \sigma_j^2\right)$.

In standard summary data MR, it is common to make the baseline assumption that there is causal homogeneity, i.e. $\beta_k = \beta$ for all k, and that all the pleiotropic effects are zero. Then we can see from (7) that there is no heterogeneity among the SNP-specific causal effects, as $\beta_j = \beta$ for all $j = 1, ..., p$. The standard error of $\widehat{\beta}_j = \widehat{\Gamma}_j / \widehat{\gamma}_j$ can be obtained using the Delta method as

$$\widehat{v}_j(\widehat{\beta}_j) = \widehat{v}_j = \sqrt{\frac{se\left(\widehat{\Gamma}_j\right)^2 + \widehat{\beta}_j^2 se\left(\widehat{\gamma}_j\right)^2}{\widehat{\gamma}_j^2}}, \tag{8}$$

but it is typically approximated by $\widehat{v}_j(\widehat{\beta}_j) = se\left(\widehat{\Gamma}_j\right)/\widehat{\gamma}_j$ since $se\left(\widehat{\gamma}_j\right)$ is deemed negligible when we only use SNPs that pass a genome-wide significance threshold as instruments [13]. This estimated standard error is then related to the analysis weight via $\widehat{v}_j = w_j^{-\frac{1}{2}}$. In this case, β can be estimated consistently by the IVW estimator as [14]:

$$\widehat{\beta}_{IVW} = \frac{\sum_{j=1}^{p} \widehat{\beta}_j w_j}{\sum_{j=1}^{p} w_j}. \tag{9}$$

The IVW estimator remains unbiased even if in fact all SNPs do exert a pleiotropic effect on the outcome, but the pleiotropy is 'balanced', that is, $E[\alpha_j] = 0$ and $Cov(\alpha_j, \gamma_j) = 0$, with the second condition being the InSIDE assumption [4].

However, standard MR may fail to provide valid causal inference if either the causal homogeneity assumption, the balanced pleiotropy assumption, or indeed both assumptions are violated by the entire set of SNPs. To this end, we make use of the definition of an instrument group as in Guo et al. [15]:

Definition 1. *A group $\mathcal{G}_s$ is a set of instruments that have the same Wald ratio estimand β_s.*

Based on this definition, SNPs belong to the same group if they identify the same causal effect, and hence have similar Wald ratio estimates. We can then re-frame the question of uncovering mechanistic heterogeneity as discovering statistical heterogeneity amongst the Wald ratio estimates. That is, if all SNPs belong to the same group, we expect to observe homogeneity among their SNP-specific causal estimates, so that Cochran's Q statistic

$$Q = \sum_{j=1}^{p} w_j (\hat{\beta}_j - \hat{\beta}_{IVW})^2 \tag{10}$$

is a reasonable draw from a χ^2 distribution on $p - 1$ degrees of freedom. If there is mechanistic heterogeneity, as revealed by a large Q statistic, this would be an unreasonable assumption. We would then seek to group SNPs into M clusters $\mathcal{S}_1 \ldots \mathcal{S}_M$ in order to define a stratified Q

statistic on $p - M$ degrees of freedom:

$$Q_{\mathcal{S}} = \sum_{m=1}^{M} Q_{\mathcal{S}_m} = \sum_{m=1}^{M} \sum_{j \in \mathcal{S}_m} w_j (\hat{\beta}_j - \hat{\beta}^{IVW}_{\mathcal{S}_m})^2, \tag{11}$$

where

- $Q_{\mathcal{S}_1} \dots Q_{\mathcal{S}_M}$ do not exhibit heterogeneity;
- Overall, $\frac{Q_{\mathcal{S}}}{p-M} \ll \frac{Q}{p-1}$.

We propose to achieve this aim by adapting the AHC method developed in Apfel and Liang [10] to the two-sample summary data MR, called 'MR-AHC', which is introduced in the next section.

3 The MR-AHC algorithm

With summary statistics $\left\{\widehat{\beta}_j, \widehat{v}_j\right\}_{j=1}^{p}$ as inputs, we apply MR-AHC to discover the SNP groups using a two-step procedure: Step 1 (agglomerative hierarchical clustering) generates a decision path from $\mathcal{K} = p$ to $\mathcal{K} = 1$ clusters; Step 2 (downward testing) re-traces the path from $\mathcal{K} = 1$ to $\mathcal{K} = p$ until the optimal cluster choice $\mathcal{K}_{opt}$ is chosen. The first step is summarized in Algorithm 1 [16]:

Algorithm 1. *Ward's algorithm*

1. ***Initialization:*** *Each SNP-specific estimate is viewed as a cluster on its own. Hence, initially, the total number of clusters is $\mathcal{K}$=p.*

2. ***Merging:*** *The two clusters that are closest as measured by their weighted squared Euclidean distance are merged into a new cluster. Without loss of generality, assume this is satisfied by cluster $\mathcal{S}_k$ and $\mathcal{S}_l$. $\widehat{\beta}^{IVW}_{\mathcal{S}_k}$ and $\widehat{\beta}^{IVW}_{\mathcal{S}_l}$ are defined as the IVW mean of all the SNP-specific estimates in $\mathcal{S}_k$ and $\mathcal{S}_l$ respectively as in (9). Further, W_k is the weight assigned to $\mathcal{S}_k$ and W_l is the weight assigned to $\mathcal{S}_l$. We then have:*

$$D_{k,l} = \frac{W_k W_l}{W_k + W_l} \left(\widehat{\beta}^{IVW}_{\mathcal{S}_k} - \widehat{\beta}^{IVW}_{\mathcal{S}_l}\right)^2 \tag{12}$$

   where $D_{k,l}$ denotes the weighted squared Euclidean distance between $\mathcal{S}_k$ and $\mathcal{S}_l$.

3. ***Iteration:*** *The merging step is repeated until all the SNP-specific estimates are in one cluster of size p.*

The weight W_k for cluster $\mathcal{S}_k$ is traditionally defined in the AHC algorithm as the number of instruments in $\mathcal{S}_k$ [16]. This can be viewed as assigning weight 1 to each Wald estimate in

the cluster so that $W_k = \sum_{j \in \mathcal{S}_k} 1 = |\mathcal{S}_k|$. MR-AHC adjusts the algorithm to take into account the uncertainty of the clustering objects, by assigning each Wald estimate its inverse-variance weight instead so that $W_k = \sum_{j \in \mathcal{S}_k} w_j$. In Appendix A we show that $D_{k,l}$ defined in (12) is essentially the Wald statistic for the null hypothesis

$$H_0 : \beta_{\mathcal{S}_k} = \beta_{\mathcal{S}_l},$$

where $\beta_{\mathcal{S}_k}$ and $\beta_{\mathcal{S}_l}$ are the cluster-specific causal estimands for $\mathcal{S}_k$ and $\mathcal{S}_l$ respectively. Therefore, $D_{k,l}$ follows a χ^2_1 distribution under the null, and merging two clusters with the smallest $D_{k,l}$ can be interpreted as merging two clusters with the highest similarity in their cluster-indicated causal effects.

The merging step is illustrated in Figure 3 (left), which shows a situation with six SNPs that form three groups (one of them comprised of a single SNP). The dotted lines at β_1 and β_2 are the true causal effects, and the circles above the real line denote the SNP-specific Wald estimates. The differences in the size of the circles reflect the fact that summary data estimates exhibit varying degrees of uncertainty. In the explanation below, we refer to these estimates and their corresponding SNPs by the numbers 1 to 6, from left to right.

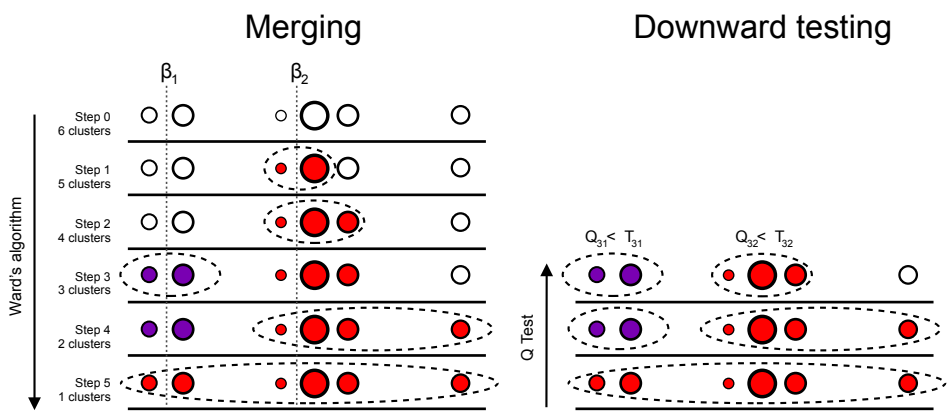


Figure 3: Illustration of MR-AHC for a hypothetical example adapted from Apfel and Liang [10].

In the initialization step of the clustering process, each SNP-specific estimate has its own cluster. Next, we merge the two estimates which are closest in terms of their weighted squared Euclidean distance, i.e. those estimated with SNPs 3 and 4 (the two red circles). These two estimates now form one cluster and we now have five clusters left. We re-calculate the distances with the five clusters and merge the closest two into a new cluster. We continue with this procedure until step 5 where all SNPs are in a single cluster.

After generating the clustering path using Algorithm 1, we are left with a $\mathcal{K} = 1$ super-cluster containing all SNPs. We then re-trace the pathway to select the optimal value of $\mathcal{K}$ using

a downward testing procedure originally proposed by Andrews [17], operates as follows:

Algorithm 2. *Downward testing procedure.*

Firstly, define Q_{fg} to be the Q statistic associated with cluster $\mathcal{S}_g$ at step f of the clustering path. Also define T_{fg} to be the $(1-\zeta)$ threshold of a χ^2 distribution on $|\mathcal{S}_{fg}|-1$ degrees of freedom.

1. *Starting from Step p-1 in Algorithm 1 where the number of clusters $\mathcal{K}$=1, calculate the global Q statistic, Q_{11};*
2. *If $Q_{11} < T_{11}$, T_{11} being a pre-specified significance threshold, then stop and assume that all the SNPs form a single cluster. If $Q_{11} \geq T_{11}$ then revert to the SNP grouping in step $p-2$ of Algorithm 1, where $\mathcal{K}$=2;*
3. *Calculate Q statistics for the two sub-clusters separately, Q_{21} and Q_{22};*
4. *If both $Q_{21} < T_{21}$ and $Q_{22} < T_{22}$ then stop. Otherwise, continue to step $p-3$ where $\mathcal{K}$=3;*
5. *Repeat steps 3-4 until a $\mathcal{K} \in (1,..,p)$ is arrived at for which no sub-cluster heterogeneity statistic rejects at its given threshold.*

Following the recommendation in Belloni et al. [18], we define the threshold for T_{fg} as $\zeta = 0.1/\log(n)$ where n is the sample size. This allows T_{fg} to increase with sample size n, but at a slower rate than n, resulting in a consistent selection procedure. That is, as n increases, the probability of correctly identifying all true members of each cluster tends to 1. See Appendix B for details. In two-sample MR, if the sample sizes are different, we recommend using the sample size for the outcome to calculate ζ. Returning to our illustrative example, we would expect the downward testing procedure to re-trace a path from step 5 to step 3 of Algorithm 1, thus determining three groups formed by SNPs 1-2, SNPs 3-5, and SNP 6 alone.

MR-AHC does not require pre-specification of the number of clusters. It can also easily identify a "null cluster" and a "junk cluster", following the terminology of Foley et al. [7] within MR-Clust, which refer to, respectively, the cluster identifying a zero causal effect, and the cluster containing SNPs not assigned to any detected clusters. Specifically, we conduct a post-clustering Wald test on each cluster-specific IVW estimate (e.g. $\widehat{\beta}^{IVW}_{\mathcal{S}_k}$ for cluster $\mathcal{S}_k$) for the null hypothesis $H_0 : \beta_{\mathcal{S}_k} = 0$. If the test statistic is smaller than the $(1-\zeta)$ threshold of a χ^2_1 distribution, then we classify $\mathcal{S}_k$ as the null cluster. For the junk cluster, we simply classify all SNPs that do not fit into any other clusters as junk SNPs.

3.1 Extending MR-AHC for summary data MR with two outcomes and a common exposure

We now extend MR-AHC to the summary data MR setting with multiple outcomes and a common exposure, which is useful for aiding the investigation of the mechanisms of multimorbidity. We focus on the common scenario of multimorbidity which involves a pair of genetically correlated LTCs, and hence extend MR-AHC to allow for two outcomes. Generalization of the method to allow for more than two outcomes follows the same principle as the two-outcome extension. In many cases, the genetic correlation between the LTCs can be attributed to them causally affected by a common risk factor [19], as illustrated by the left panel of Figure 4 where r_g is the genetic correlation between the two LTCs. For example, consider the multimorbid LTC pair T2D and OA, MR studies have shown strong evidence that BMI, as a proxy for obesity, leads to increasing risks for both conditions [20]. In such cases, to discover the shared genetic pathway between the two conditions, a natural strategy is to investigate the subsets of the SNPs associated with the common exposure that may indicate distinct pathways from the common exposure to the LTC pair.

A hypothetical example for this is presented in the right panel of Figure 4, where the SNPs associated with the exposure X are split into three groups (G_1 to G_3), as they affect the two outcomes Y_1 and Y_2 through three different aspects of the exposure (X_1 to X_3). Among the three SNP groups, G_1 indicates positive effect on Y_1 but negative effect on Y_2, and G_3 indicates positive effect on Y_1 and no effect on Y_2. Only the group G_2 is associated with increasing risks of both Y_1 and Y_2. This motivates the aforementioned extension of MR-AHC, which classifies the SNPs into distinct clusters based on their SNP-specific causal estimates for both outcomes.

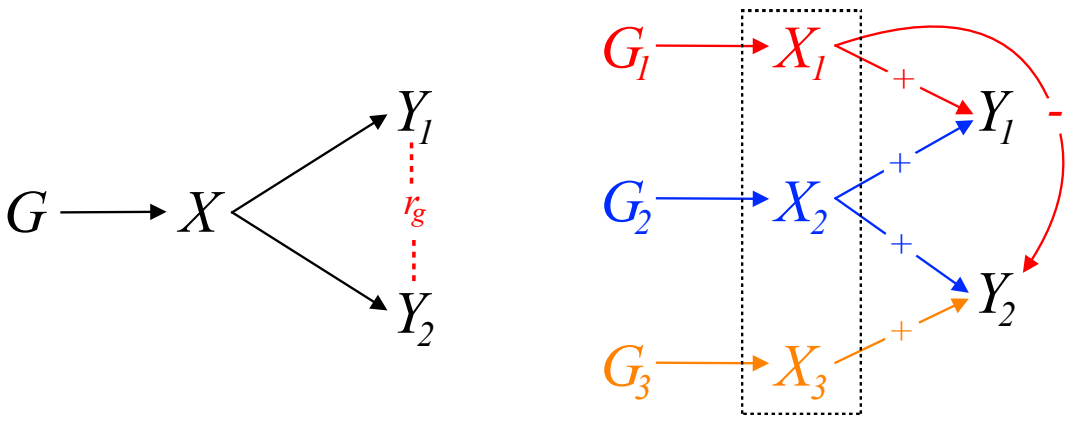


Figure 4: Left: Genetic correlation between two traits due to a common risk factor; right: causal effect heterogeneity in the exposure with two outcomes.

Similar to the previous setup with one outcome and one exposure, we assume that Model (1)-(4) holds for both outcomes, that is, Y_1 and Y_2. Let $\left\{\widehat{\Gamma}_{1j}, se\left(\widehat{\Gamma}_{1j}\right)\right\}_{j=1}^{p}$ and $\left\{\widehat{\Gamma}_{2j}, se\left(\widehat{\Gamma}_{2j}\right)\right\}_{j=1}^{p}$ denote the summary statistics for the $G-Y_1$ and $G-Y_2$ relationship respectively. We assume that these statistics are obtained from cohorts that are independent of the exposure sample from which we obtain $\left\{\widehat{\gamma}_j, se\left(\widehat{\gamma}_j\right)\right\}_{j=1}^{p}$, the $G-X$ summary statistics, but we allow for overlap of the Y_1 and Y_2 samples. For a single SNP G_j, let $\widehat{\beta}_j$ be its Wald ratio estimate for outcome Y_1, and $\widehat{\theta}_j$ for Y_2. The extension of Algorithm 1 and 2 can be summarized as extending two quantities to allow for both $\left(\widehat{\beta}_j, \widehat{\theta}_j\right)$: the weighted squared Euclidean distance defined in (12), and the Q statistic defined in (10). Let the covariance between the two SNP-specific estimates be $Cov(\widehat{\beta}_j, \widehat{\theta}_j) = \rho$ and assume that ρ is the same across all $j = 1, ..., p$. We show that ρ is related to the phenotypic covariance $Cov(Y_1, Y_2)$ and can be estimated with summary statistics using LD score regression [21]. When the $G-Y_1$ and $G-Y_2$ summary statistics are estimated with non-overlapping samples, we have $\rho = 0$, see Appendix C for details. Let S_k and S_l be two arbitrary non-empty SNP clusters. Let $\widehat{\beta}_{S_k}^{IVW}$ be the IVW mean of all the SNP-specific estimates within S_k for outcome Y_1, and $\widehat{\theta}_{S_k}^{IVW}$ for outcome Y_2:

$$\widehat{\beta}_{S_k}^{IVW} = \frac{\sum_{j=1}^{|S_k|} \widehat{\beta}_j w_{1j}}{\sum_{j=1}^{|S_k|} w_{1j}} \quad \text{and} \quad \widehat{\theta}_{S_k}^{IVW} = \frac{\sum_{j=1}^{|S_k|} \widehat{\theta}_j w_{2j}}{\sum_{j=1}^{|S_k|} w_{2j}},$$

where the weights $w_{1j} = 1/se(\widehat{\beta}_j)^2$ and $w_{2j} = 1/se(\widehat{\theta}_j)^2$. $|S_k|$ is the number of SNPs within S_k. Let $W_{1k} = \sum_{j=1}^{|S_k|} w_{1j}$, and $W_{2k} = \sum_{j=1}^{|S_k|} w_{2j}$. Then the extended weighted squared Euclidean distance is (see Appendix C for the derivation):

$$\mathcal{D}_{k,l} = (\widehat{\beta}_{S_k}^{IVW} - \widehat{\beta}_{S_l}^{IVW}, \widehat{\theta}_{S_k}^{IVW} - \widehat{\theta}_{S_l}^{IVW}) \begin{bmatrix} \frac{W_{1k}+W_{1l}}{W_{1k}W_{1l}} & \rho_{kl} \\ \rho_{kl} & \frac{W_{2k}+W_{2l}}{W_{2k}W_{2l}} \end{bmatrix} \begin{pmatrix} \widehat{\beta}_{S_k}^{IVW} - \widehat{\beta}_{S_l}^{IVW} \\ \widehat{\theta}_{S_k}^{IVW} - \widehat{\theta}_{S_l}^{IVW} \end{pmatrix}, \tag{13}$$

where $\rho_{kl} = \frac{\rho}{W_{1k}W_{2k}} \sum_{j=1}^{|S_k|} w_{1j}w_{2j} + \frac{\rho}{W_{1l}W_{2l}} \sum_{j=1}^{|S_l|} w_{1j}w_{2j}$. Let $\widehat{\boldsymbol{\beta}}_k$ be the $|S_k|$-vector of all SNP-specific estimates for Y_1 in S_k: $\widehat{\boldsymbol{\beta}}_k = \left(\widehat{\beta}_1, ..., \widehat{\beta}_{|S_k|}\right)'$. Similarly, define $\widehat{\boldsymbol{\theta}}_k$ for Y_2: $\widehat{\boldsymbol{\theta}}_k = \left(\widehat{\theta}_1, ..., \widehat{\theta}_{|S_k|}\right)'$. Then the extended Q statistic for S_k is

$$Q_k = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\iota}_k \widehat{\beta}_{S_k}^{IVW} \\ \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\iota}_k \widehat{\theta}_{S_k}^{IVW} \end{pmatrix}' \begin{bmatrix} Var(\widehat{\boldsymbol{\beta}}_k) & \rho\mathbf{I} \\ \rho\mathbf{I} & Var(\widehat{\boldsymbol{\theta}}_k) \end{bmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\iota}_k \widehat{\beta}_{S_k}^{IVW} \\ \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\iota}_k \widehat{\theta}_{S_k}^{IVW} \end{pmatrix}, \tag{14}$$

where $\boldsymbol{\iota}_k$ is a $|S_k|$-vector of 1. $Var(\widehat{\boldsymbol{\beta}}_k)$ is a $|S_k| \times |S_k|$ diagonal matrix with entries $\left(1/se(\widehat{\beta}_1)^2, ..., 1/se(\widehat{\beta}_{|S_k|})^2\right)$, and $Var(\widehat{\boldsymbol{\theta}}_k)$ with entries $\left(1/se(\widehat{\theta}_1)^2, ..., 1/se(\widehat{\theta}_{|S_k|})^2\right)$. $\mathbf{I}$ is the $|S_k| \times |S_k|$ identity matrix. Q_k follows a χ^2 distribution on $2(|S_k| - 1)$ degrees of freedom.

4 Implementation

We conduct Monte Carlo simulations to evaluate the performance of MR-AHC, and to compare with MR-Clust [7], an popular method for performing cluster analysis in MR. MR-Clust assumes a normal mixture distribution model for the SNP-specific causal estimates parameterised by cluster means, variances and cluster membership proportions. The Expectation-Maximization (EM) algorithm is used for estimation. It is a 'soft clustering' method in the sense that it usually assigns a SNP to multiple clusters with different probabilities. The MR-AHC method on the other hand is 'hard clustering', where each SNP is assigned to exactly one cluster.

In implementing the MR-AHC method, in addition to the baseline procedure summarized in Algorithm 1 and 2, we propose a two-step extension of the method to handle outliers in the Wald estimates: after we run Algorithm 1 and 2 and obtain the clustering results, within each detected cluster, we calculate each individual SNP's contribution to the overall Q statistic of that cluster. The individual Q statistic approximately follows a χ_1^2 distribution [13], and SNPs with large individual Q (here defined as the p-value of the individual Q below 5%) are viewed as outliers. We remove the outliers from each detected cluster, and re-run Algorithm 1 and 2 with all the remaining SNPs. All the outliers are classified as junk SNPs.

In simulation studies, we report the MR-AHC results of both the baseline procedure (as Version A) and the two-step outlier removal extension (as Version B). For MR-Clust, following the recommendation in Foley et al. [7], we also report two versions: Version A is based on the 'best' clustering results, i.e. assigning each SNP to the cluster with the highest probability. For Version B, a SNP is only assigned to a cluster if its highest probability is no less than 0.8. In an applied example where we investigate the causal effects of adiposity on a pair of LTCs, T2D and OA, we also implement the methods in both versions where feasible.

4.1 Simulation study

We conduct two sets of simulations to evaluate the performance of the methods in detecting the variant clusters and estimating the causal effects in various MR settings. First we compare the performance of MR-AHC and MR-Clust in the setup with one outcome and one exposure. Then we test MR-AHC in simulations with two outcomes and a common exposure. See Figure 5 for an illustration of the main simulation designs, with additional ones shown in Figure S1 in Appendix D.2. We also examine the computational efficiency of the two methods as measured by the average running time when varying the number of SNPs involved in the analysis.

Overall, the performance of MR-AHC is highly comparable with MR-Clust, and the clustering results of both methods are close to the ground truth. In the two-outcome designs,

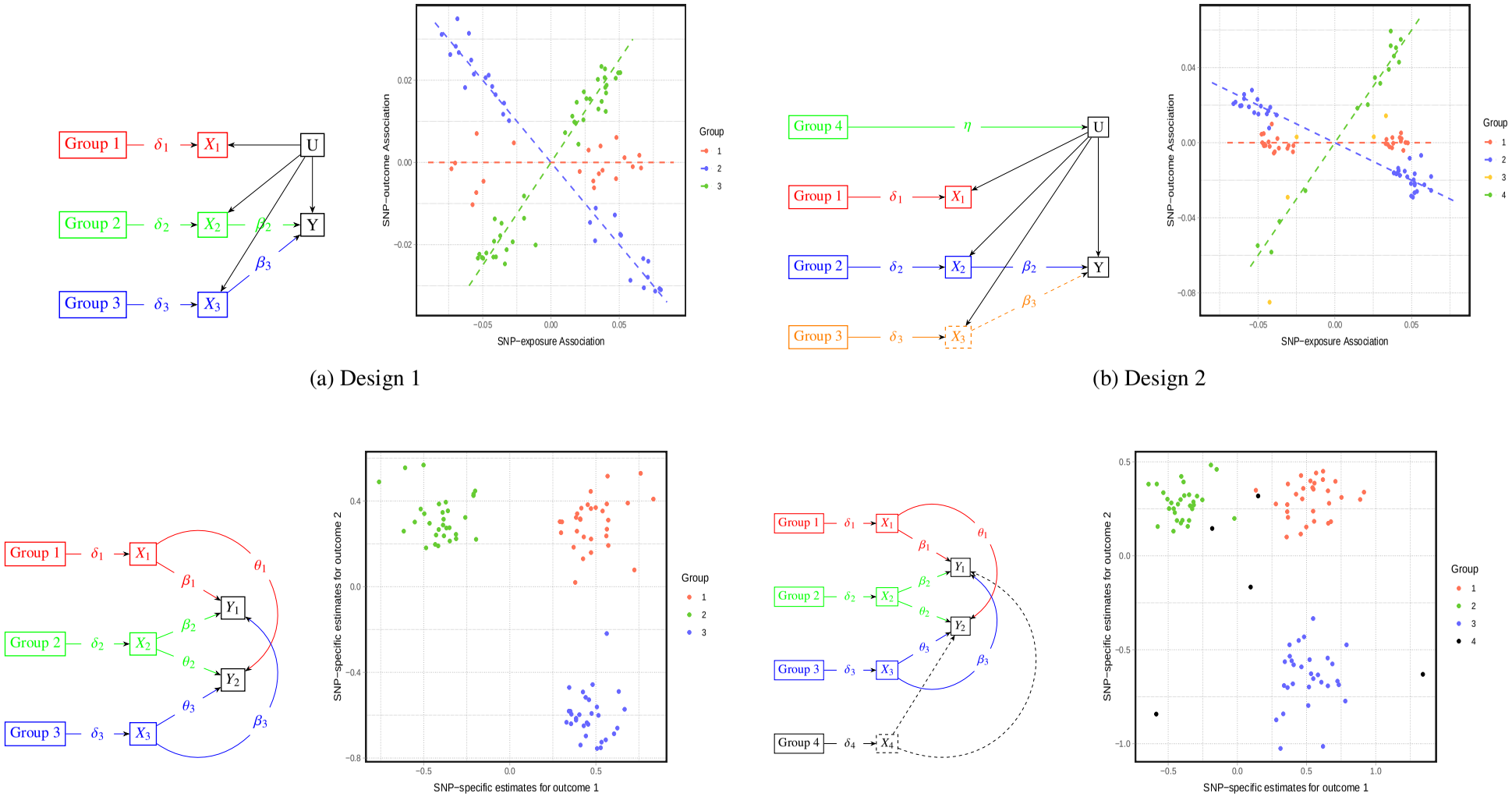
MR-AHC also performs well. For MR-AHC, Version B outperforms the baseline Version A, therefore we recommend the outlier removal procedure in the implementation of our method.

In terms of computational efficiency, MR-AHC is much faster than MR-Clust, as illustrated in Figure 6. We show their absolute running time in seconds (left panel of Figure 6), and the ratio of MR-Clust running time to MR-AHC running time (right panel of Figure 6). The results suggest that for the majority of MR studies which typically involve an order of 200-300 genome-wide significant genetic variants, MR-AHC has a significant computational advantage. This is especially relevant in phenome-wide studies where we carry out clustering analysis for a large number of traits across the human phenome [22]. For example, on the same hardware configuration as we use to test the running time, if conducting clustering analysis at scale for 1000 traits, with each study involving 200 SNPs on average, the running time of MR-AHC would be $\approx$12 minutes, while for MR-Clust it would be $\approx$10 hours. See Appendix D for details of the simulation study.

4.2 Application to estimating the causal effects of adiposity on type 2 diabetes and osteoarthritis

We apply our method to investigate the causal relationship between body fat percentage (BFP), as a common exposure, for two outcomes T2D and OA using a three-sample summary data MR design. Specifically, we use SNP-BFP summary data from a GWAS study based on UK Biobank from Martin et al. [20]. This includes 696 BFP SNPs in total, among which 36 have been hypothesised to be favourable adiposity (FA) SNPs and 38 unfavourable adiposity (UFA) SNPs, due to their associations with various metabolic traits. This means that FA SNPs are typically associated with a decrease in metabolic risk whereas UFA SNPs are typically associated with an increase in metabolic risk [20]. The T2D GWAS data are from Mahajan et al. [23], which combine 31 published GWAS studies but excluding the UK Biobank individuals. The SNP-OA summary statistics are from a FinnGen GWAS (code: M13_ARTHROSIS_INCLAVO) [24]. Only SNPs present in all three datasets are used for MR analyses (487 in total with 32 FA and 22 UFA SNPs). The overall instrument strength measured by the F-statistic is $F = 52.57$. The T2D and OA samples are non-overlapping, therefore for each SNP, the covariance between the SNP-T2D estimate and SNP-OA estimate can be treated as zero.

We conduct two sets of analyses: first we do cluster analysis in the one-outcome MR setting on the BFP-T2D and BFP-OA relationship separately using both MR-AHC and MR-Clust. In our implementation, no "FA" or "UFA" labelling information is used, but we are nevertheless interested in whether both approaches can uncover clusters closely matching the FA and UFA groups without this external knowledge. Second, we implement the extended MR-AHC in the two-outcome MR setting to account for causal effects of BFP on both conditions simultaneously.



(a) Design 1

(b) Design 2

(c) Design 6

(d) Design 7

Figure 5: DAGs and scatter plots of the simulation designs 1, 2, 6 and 7. The scatter plots (on the right in each panel) are generated from the simulated data. For scatter plots in Panel (a) and (b), on the x-axis we have the SNP-exposure associations and y-axis the SNP-outcome associations. The slopes of the dashed lines through the origin are the true causal effects indicated by each SNP group. For scatter plots in Panel (c) and (d), on the x-axis we have the Wald estimate of each SNP for the first outcome, and on the y-axis the Wald estimate for the second outcome.

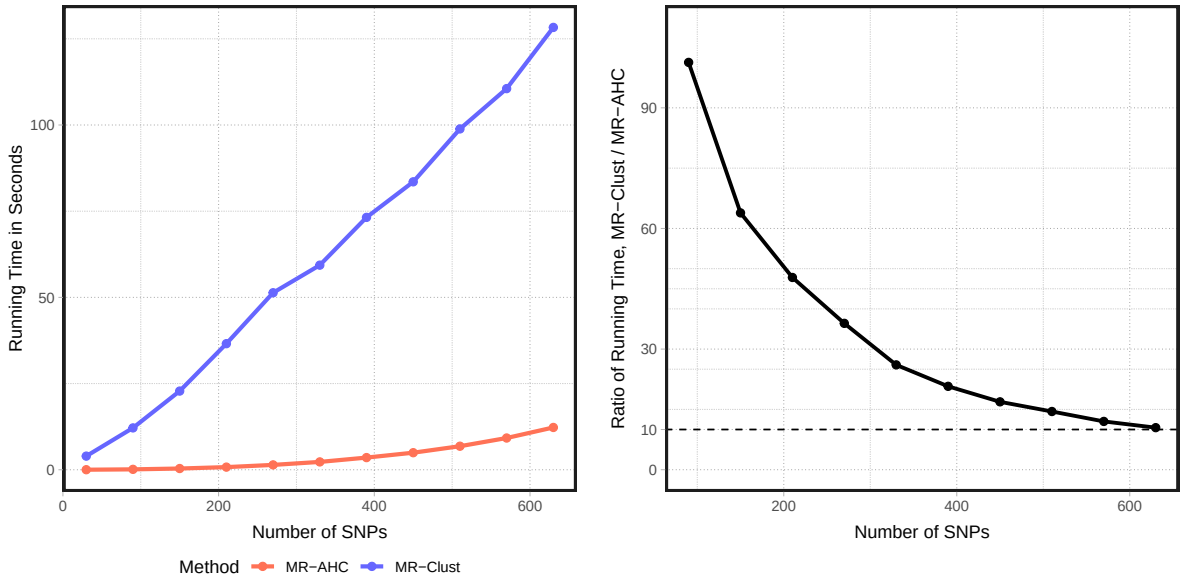


Figure 6: Running time comparison between MR-AHC and MR-Clust.

We use the effective sample size [25] to calculate the threshold p-value $0.1/\log(n)$ for MR-AHC in the binary outcome setting. Specifically, we use $n = 193,440$ for the BFP-T2D and two-outcome analysis, and $n = 254,049$ for the BFP-OA analysis. For MR-Clust, we conduct both Version A and B as described in the simulation section. For MR-AHC, in addition to Version A, we also use an iterated outlier removal procedure (Version C): this performs the outlier removal and re-fitting of Version B indefinitely, until the individual p-values of the Q statistics for all SNPs are above 5%.

4.2.1 Clustering results

Clustering results for the BFP-T2D and BFP-OA analysis are illustrated in Figure 7 using MR-AHC and MR-Clust Version A respectively. Overall, the MR-AHC and MR-Clust results share a similarity of 83.52% for BFP-T2D, and 94.64% for BFP-OA, as measured by the Rand index. For T2D, MR-AHC detects three substantive clusters and a junk cluster. Cluster 1 identifies a positive causal effect, suggesting that increasing BFP via the cluster-specific mechanism leads to increasing risk of T2D, while for clusters 2 and 3 the opposing risk-reducing effects hold. MR-Clust detects three substantive clusters plus a null cluster, displaying similar patterns as those by MR-AHC: two clusters indicate positive effects and one indicates the opposite. The clustering outcomes of both methods further support the finding that the metabolic effects of obesity on T2D are heterogeneous [20]. For OA, the causal effects indicated by the detected clusters do not display such heterogeneity: both methods only detect a cluster identifying a positive effect and a null cluster indicating zero effect.

Specifically, we check the clustering results for the pre-labeled FA and UFA SNPs. With T2D as the outcome, for MR-AHC, based on the outlier-removal results (Version C) and ex-

cluding the ones assigned to the junk cluster, 100% of the FA SNPs are assigned to clusters with negative estimates, and 78.57% of UFA SNPs are assigned to the cluster with a positive estimate. For MR-Clust (Version B), the numbers are 65.21% for FA and 77.78% for UFA SNPs. With OA as the outcome, for MR-AHC, 25.81% of the FA SNPs and 46.67% of the UFA SNPs are classified into the positive cluster. For MR-Clust it is 32% of the FA SNPs and 45% of the UFA SNPs. These results are consistent with the conclusions in Martin et al. [20] that the effect of higher adiposity on T2D is primarily metabolic as FA and UFA SNPs indicate opposing effects, while for OA the effect is mainly non-metabolic as both FA and UFA SNPs may indicate increasing risk. For detailed clustering results, see Appendix E.

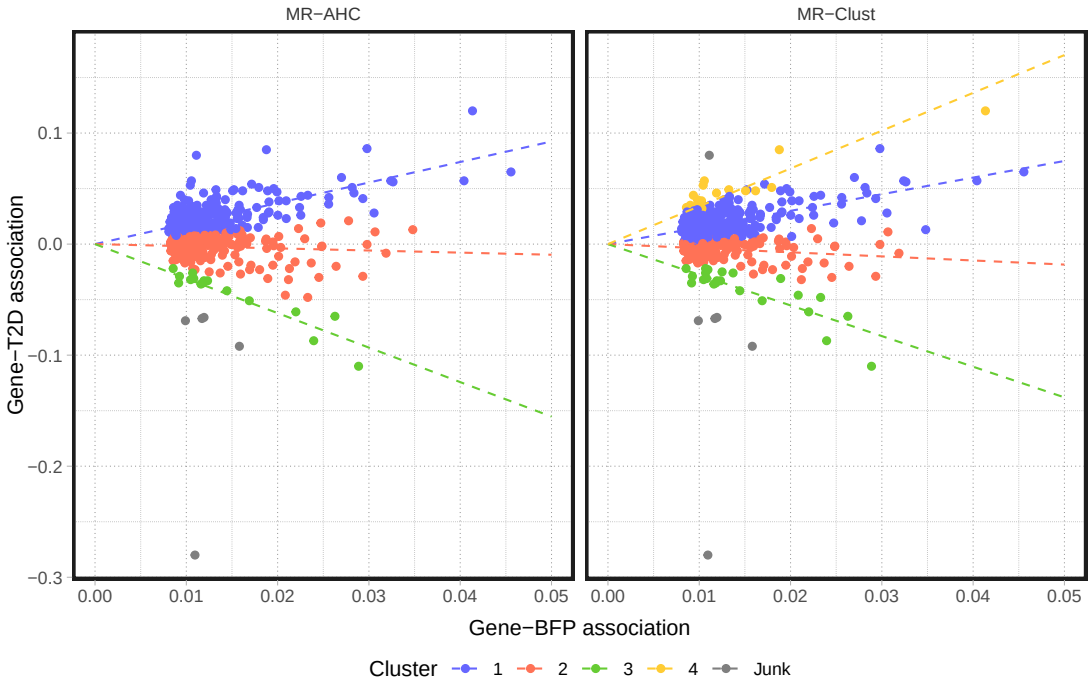
The clustering results for the pair of conditions T2D and OA as outcomes are illustrated in Figure 8. MR-AHC detects 5 substantive clusters with Version A and 4 clusters with the outlier removal procedure Version C. We base our analysis on the latter. The estimation results are presented in Table 1. The cluster-specific estimates and 95% confidence intervals in odds ratio are illustrated in Figure 9. Among the 4 clusters, it is in line with the one-outcome analysis that the clusters identify either positive or negative effects for T2D, and either positive or null effects for OA. Only Cluster 1 with 124 SNPs indicates increasing risks for both conditions. We further investigate the heterogeneous causal pathways reflected by the SNP clusters in Section 4.2.2.

Cluster	# SNPs	Est-T2D	t-T2D	Est-OA	t-OA	Q p-value	I^2
1	124	1.326 (0.073)	18.236	1.250 (0.054)	23.321	0.929	0.000
2	258	1.076 (0.051)	21.102	0.067 (0.037)	1.791	0.951	0.000
3	32	−1.068 (0.138)	7.718	0.738 (0.101)	7.327	1.000	0.000
4	22	−2.799 (0.190)	14.704	0.103 (0.143)	0.722	0.437	0.019

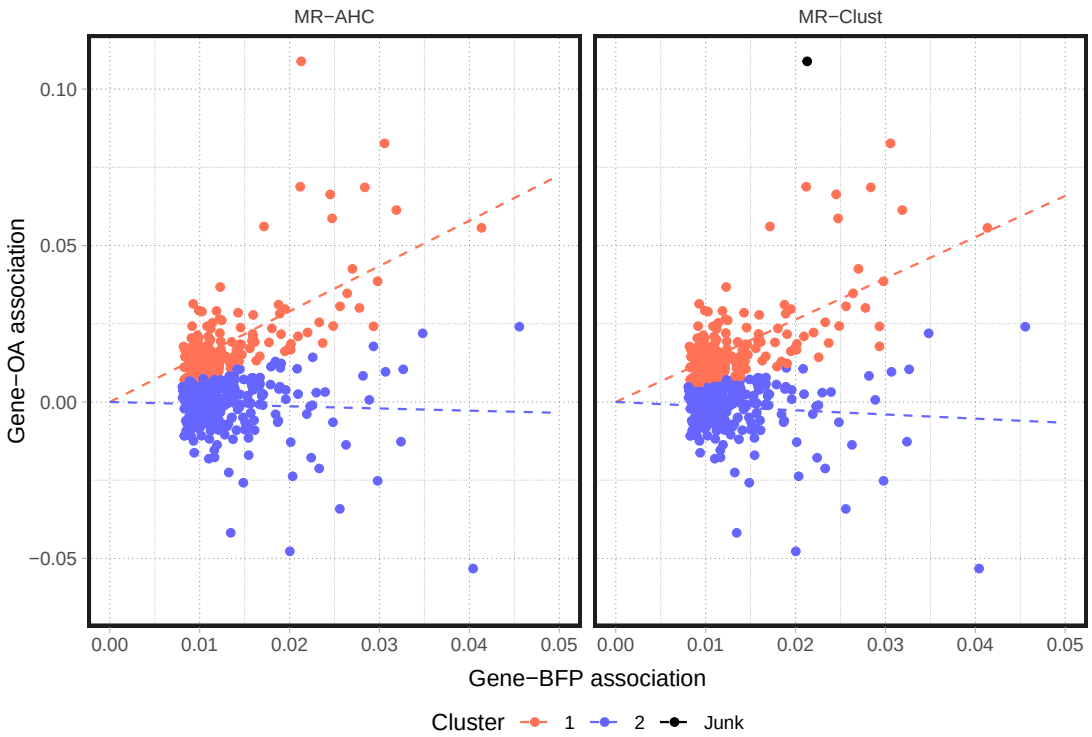
Table 1: The estimation results for BFP and T2D-OA with outlier removal. We report the number of SNPs in the cluster (*#SNPs*), the cluster-specific IVW estimate (*Est*), the standard error (in the parentheses under the estimates) and associated t-statistic (t), and statistics measuring the heterogeneity within each cluster (p-value of the Q statistic and the I^2 statistic [26]).

4.2.2 Further biological insights

To better understand the underlying biological mechanisms reflected by the clusters identified by MR-AHC in the two-outcome analysis, we use PhenoScanner [27, 28] to perform an enrichment analysis and report traits that are significantly enriched in each cluster (details in Appendix F). We then classify the traits into three main categories: obesity-related traits (body fat and body



(a) Clustering outcomes for the BFP-T2D relationship



(b) Clustering outcomes for the BFP-OA relationship

Figure 7: The MR-AHC results (left) and MR-Clust results (right). On the x-axis are the SNP-BFP associations and y-axis the SNP-outcome associations. Each point represents a specific SNP. The slopes of the dashed lines through the origin are the causal estimates identified by each SNP cluster.

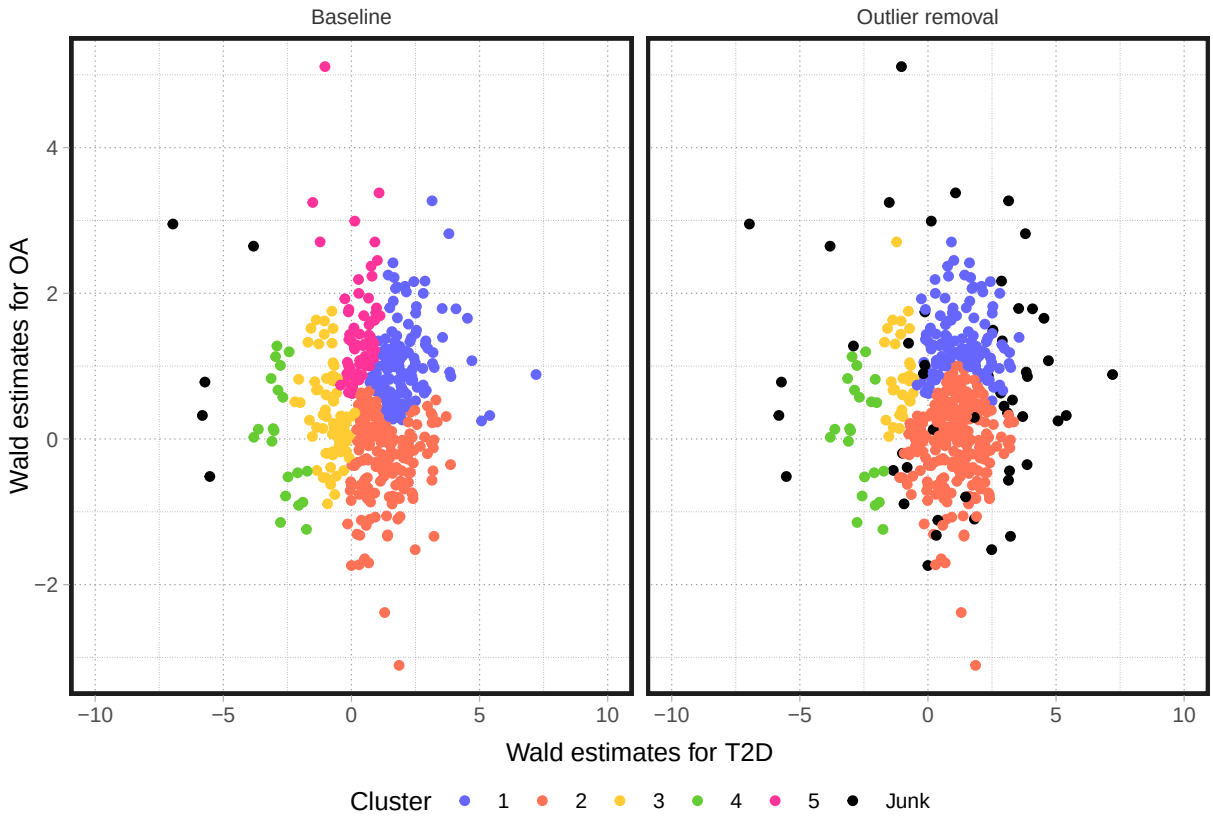


Figure 8: MR-AHC Clustering results for the causal effects of BPF on T2D and OA with the baseline algorithm (left) and the outlier-removal procedure (right). On the x-axis are the Wald estimates for T2D, and y-axis the Wald estimates for OA. Each point represents a specific SNP.

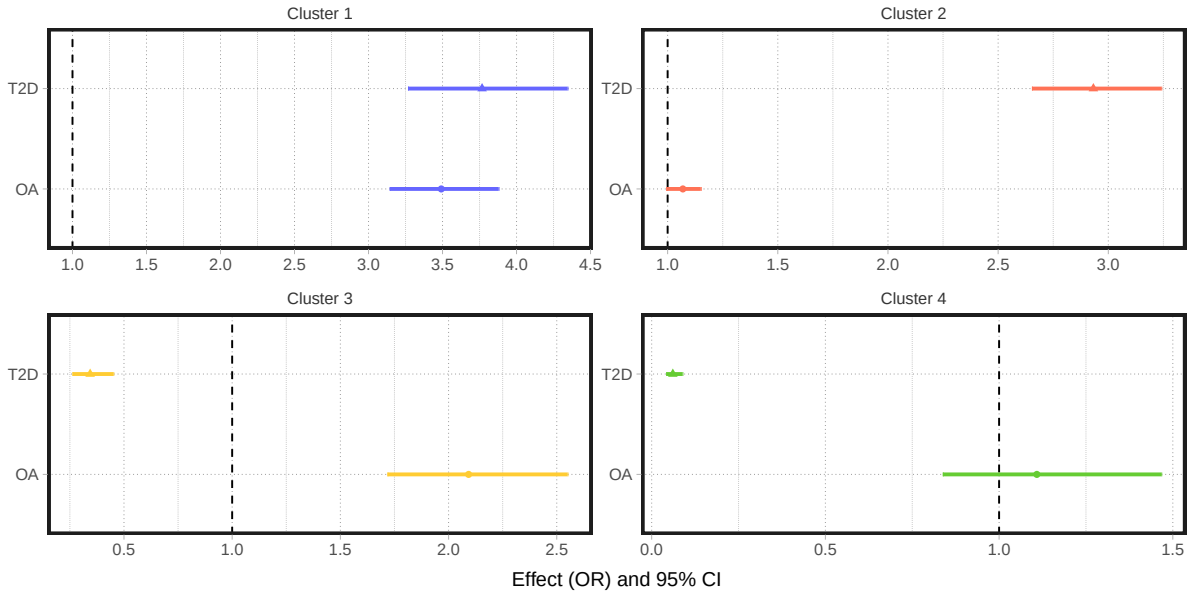


Figure 9: The cluster-specific causal estimates and 95% confidence intervals in odds ratio for each cluster detected by MR-AHC Version C.

size), cardio-metabolic traits and risk factors (lipids, blood pressure, cardiovascular diseases) and miscellaneous, see Table S24 in Appendix F. In cluster 1 (increased risk for both T2D and OA), we observe enrichment for several obesity-related traits, as well as some traits that are likely to capture lower socio-economic status (absence of qualification, more time spent watching TV, see Figure S2 A). No enriched traits were identified for cluster 2. Cluster 3 (decreased risk for T2D but increased risk for OA) is only enriched for 'usual walking pace'. Interestingly, SNPs in cluster 3 are associated with a reduction in walking pace (Figure S2 B), consistent with a slower walking speed already reported for people with OA [29]. Cluster 4 (decreased risk for T2D but no effect on OA) is the only cluster enriched for cardio-metabolic traits. More specifically, SNPs in cluster 4 are associated with lower family history of diabetes, and reduced self-reported cholesterol medication (Figure S2 C).

5 Discussion

In this paper we adapt the general method of agglomerative hierarchical clustering to the two-sample summary data Mendelian randomization setup, called 'MR-AHC'. This method is useful for interrogating a set of genetic variants to see if they collectively identify a single causal effect, or if it is more plausible that a number of subgroups identify distinct effects driven by different biological mechanisms. The method is of particular interest when the potential sub-components of the exposure are not known beforehand, or difficult/expensive to measure.

While for hierarchical clustering it can be difficult to choose the 'optimal' dissimilarity metric, linkage and number of clusters on the dendrogram to yield reliable clustering results, studies in the field of model selection [10, 17] provide theoretical results that allow us to guarantee that our method achieves accurate clustering. By using a simple but statistically efficient criterion for merging clusters, we adapt the original AHC method in Apfel and Liang [10] to further account for the varying degrees of uncertainty exhibited in summary data estimates due to varying allele frequency. Also, our method is invariant to the ordering of the input data, and can handle outliers in the variant-specific estimates with the outlier removal procedure.

Like the existing method MR-Clust [7], our method possesses the features that it does not require pre-specifying the number of clusters, and that alongside detecting meaningful clusters it can also identify and label null and junk clusters. We show in simulations that MR-AHC generates comparable results with MR-Clust with only a slight decrease in classification accuracy, but it is far more computationally efficient. From this viewpoint it can be viewed as a reliable alternative to MR-Clust and the computational efficiency of MR-AHC means that it serves as an excellent framework to apply at scale across the human phenome. We also extend MR-AHC to handle two outcomes and a common exposure, which can be used to investigate the causal mechanisms of multimorbidity. We apply our method to analyze the causal effects of body fat

percentage on a pair of multimorbid conditions, type 2 diabetes and osteoarthritis, showcasing how our method can detect variant clusters reflecting the heterogeneous shared causal pathways.

For future work, a potential extension is to adapt the method to more complex analysis challenges, such as allowing direct causality between the multimorbid conditions. Other topics for future research include the adaptation of the method to incorporate *a priori* biological knowledge as well as correction terms for weak instruments and Winner's curse.

Funding

This work was supported by the Strategic Priority Fund "Tackling multimorbidity at scale" programme (grant number MC/MR/WO14548/1) delivered by the Medical Research Council and the National Institute for Health and Care Research in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council. Nicolas Apfel is supported by the ESRC grant EST013567/1.

Data availability

For the GWAS data used in the applied examples, the BFP summary data are from the original study (Supplimentary material 1d); the T2D summary data are available from `http://diagram-consortium.org/downloads.html`; the OA summary data are available from `https://r8.risteys.finngen.fi/phenocode/M13_ARTHROSIS_INCLAVO`. The R code that implements MR-AHC and that generates the simulation datasets are available on Github: `https://github.com/xiaoran-liang/MRAHC`.

References

[1] George Davey Smith and Shah Ebrahim. "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" In: *International Journal of Epidemiology* 32.1 (Feb. 2003), pp. 1–22. ISSN: 0300-5771. DOI: 10.1093/ije/dyg070. URL: https://doi.org/10.1093/ije/dyg070.

[2] Jack Bowden and Michael V Holmes. "Meta-analysis and Mendelian randomization: A review". In: *Research synthesis methods* 10.4 (2019), pp. 486–496.

[3] Jack Bowden and Michael V. Holmes. "Meta-analysis and Mendelian randomization: A review". In: *Research Synthesis Methods* 10.4 (2019), pp. 486–496. DOI: https://doi.org/10.1002/jrsm.1346. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1346.

[4] Jack Bowden, George Davey Smith, and Stephen Burgess. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression". In: *International journal of epidemiology* 44.2 (2015), pp. 512–525.

[5] Gibran Hemani, Jack Bowden, and George Davey Smith. "Evaluating the potential role of pleiotropy in Mendelian randomization studies". In: *Human Molecular Genetics* 27.R2 (May 2018), R195–R208. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy163. URL: https://doi.org/10.1093/hmg/ddy163.

[6] Chin Yang Shapland, Qingyuan Zhao, and Jack Bowden. "Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy". In: *Statistics in Medicine* 41.6 (2022), pp. 1100–1119. DOI: https://doi.org/10.1002/sim.9320. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9320.

[7] Christopher N Foley et al. "MR-Clust: clustering of genetic variants in Mendelian randomization with similar causal estimates". In: *Bioinformatics* 37.4 (2021), pp. 531–541.

[8] Daniel Iong, Qingyuan Zhao, and Yang Chen. "A latent mixture model for heterogeneous causal mechanisms in Mendelian randomization". In: *arXiv preprint arXiv:2007.06476* (2020).

[9] Susan Martin et al. "Genetic evidence for different adiposity phenotypes and their opposing influences on ectopic fat and risk of cardiometabolic disease". In: *Diabetes* 70.8 (2021), pp. 1843–1856.

[10] Nicolas Apfel and Xiaoran Liang. "Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables". In: *arXiv preprint arXiv:2101.05774* (2021).

[11] Alessandra Marengoni et al. "Aging with multimorbidity: a systematic review of the literature". In: *Ageing research reviews* 10.4 (2011), pp. 430–439.

[12] Guiying Dong et al. “A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank”. In: *Genome medicine* 13.1 (2021), pp. 1–20.

[13] Jack Bowden et al. “Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption”. In: *International Journal of Epidemiology* 48.3 (Dec. 2018), pp. 728–742. ISSN: 0300-5771. DOI: 10.1093/ije/dyy258. URL: https://doi.org/10.1093/ije/dyy258.

[14] Stephen Burgess, Adam Butterworth, and Simon G Thompson. “Mendelian randomization analysis with multiple genetic variants using summarized data”. In: *Genetic epidemiology* 37.7 (2013), pp. 658–665.

[15] Zijian Guo et al. “Confidence Intervals for Causal Effects with Invalid Instruments by Using Two-Stage Hard Thresholding with Voting”. In: *Journal of the Royal Statistical Society: Series B* 80.4 (2018), pp. 793–815. DOI: 10.1111/rssb.12275.

[16] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[17] Donald W. K. Andrews. “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation”. In: *Econometrica* 67.3 (1999), pp. 543–563. DOI: 10.1111/1468-0262.00036.

[18] A. Belloni et al. “Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain”. In: *Econometrica* 80.6 (2012), pp. 2369–2429. DOI: 10.3982/ecta9626.

[19] Wouter Van Rheenen et al. “Genetic correlations of polygenic disease traits: from theory to practice”. In: *Nature Reviews Genetics* 20.10 (2019), pp. 567–581.

[20] Susan Martin et al. “Disease consequences of higher adiposity uncoupled from its adverse metabolic effects using Mendelian randomisation”. In: *Elife* 11 (2022), e72452.

[21] Brendan Bulik-Sullivan et al. “An atlas of genetic correlations across human diseases and traits”. In: *Nature genetics* 47.11 (2015), pp. 1236–1241.

[22] Lingyan Chen et al. “Systematic Mendelian randomization using the human plasma proteome to discover potential therapeutic targets for stroke”. en. In: *Nat. Commun.* 13.1 (Oct. 2022), p. 6143.

[23] Anubha Mahajan et al. “Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps”. In: *Nature genetics* 50.11 (2018), pp. 1505–1513.

[24] FinnGen. *FinnGen Documentation of R8 release*. 2022. URL: https://finngen.gitbook.io/documentation/.

[25] Buhm Han and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies". In: *The American Journal of Human Genetics* 88.5 (2011), pp. 586–598.

[26] Julian PT Higgins and Simon G Thompson. "Quantifying heterogeneity in a meta-analysis". In: *Statistics in medicine* 21.11 (2002), pp. 1539–1558.

[27] James R. Staley et al. "PhenoScanner: a database of human genotype–phenotype associations". In: *Bioinformatics* 32.20 (June 2016), pp. 3207–3209. DOI: 10.1093/bioinformatics/btw373. URL: https://doi.org/10.1093/bioinformatics/btw373.

[28] Mihir A Kamat et al. "PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations". In: *Bioinformatics* 35.22 (June 2019). Ed. by Janet Kelso, pp. 4851–4853. DOI: 10.1093/bioinformatics/btz469. URL: https://doi.org/10.1093/bioinformatics/btz469.

[29] Jama L. Purser et al. "Slower walking speed is associated with incident knee osteoarthritis-related outcomes". In: *Arthritis Care &amp Research* (2012), n/a–n/a. DOI: 10.1002/acr.21655. URL: https://doi.org/10.1002/acr.21655.

[30] Fabiola Del Greco M et al. "Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome". In: *Statistics in medicine* 34.21 (2015), pp. 2926–2940.

[31] Jingshu Wang et al. "Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments". In: *PLoS genetics* 17.6 (2021), e1009575.

[32] William M Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.

[33] Qingyuan Zhao et al. "Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score". In: *The Annals of Statistics* 48.3 (2020), pp. 1742–1769.

[34] Lucas Parra and Harrison H Barrett. "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET". In: *IEEE transactions on medical imaging* 17.2 (1998), pp. 228–235.

[35] Renato Cordeiro de Amorim, Vladimir Makarenkov, and Boris Mirkin. "A-Wardpβ: Effective hierarchical clustering using the Minkowski metric and a fast k-means initialisation". In: *Information Sciences* 370 (2016), pp. 343–354.

[36] Lei Xu and Michael I Jordan. "On convergence properties of the EM algorithm for Gaussian mixtures". In: *Neural computation* 8.1 (1996), pp. 129–151.