Received: 30 September 2021 | Revised: 28 March 2022 | Accepted: 25 April 2022

DOI: 10.1002/gepi.22462

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

WILEY

Disentangling the effects of traits with shared clustered genetic predictors using multivariable Mendelian randomization

Fatima Batool[1] | Ashish Patel[1] | Dipender Gill[2,3,4] | Stephen Burgess[1,5]

[1]MRC Biostatistics Unit, Institute of Public Health, Biomedical Campus, University of Cambridge, Cambridge, UK

[2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

[3]Genetics Department, Novo Nordisk Research Centre, Oxford, UK

[4]Clinical Pharmacology and Therapeutics Section, Institute for Infection and Immunity, St George's, University of London, London, UK

[5]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

Correspondence
Stephen Burgess, MRC Biostatistics Unit, Institute of Public Health, Biomedical Campus, University of Cambridge, Cambridge, UK.
Email: sb452@medschl.cam.ac.uk and stephen.burgess@mrc-bsu.cam.ac.uk

Funding information
Sir Henry Dale Fellowships, Grant/Award Number: 204623/Z/16/Z; UK Research and Innovation, Grant/Award Numbers: MC_UU_00002_7, MR/S037675/1; British Heart Foundation, Grant/Award Number: RE/18/4/34215; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014

Abstract

When genetic variants in a gene cluster are associated with a disease outcome, the causal pathway from the variants to the outcome can be difficult to disentangle. For example, the chemokine receptor gene cluster contains genetic variants associated with various cytokines. Associations between variants in this cluster and stroke risk may be driven by any of these cytokines. Multivariable Mendelian randomization is an extension of standard univariable Mendelian randomization to estimate the direct effects of related exposures with shared genetic predictors. However, when genetic variants are clustered, due to being located in a single genetic region, a Goldilocks dilemma arises: including too many highly-correlated variants in the analysis can lead to ill-conditioning, but pruning variants too aggressively can lead to imprecise estimates or even lack of identification. We propose multivariable methods that use principal component analysis to reduce many correlated genetic variants into a smaller number of orthogonal components that are used as instrumental variables. We show in simulations that these methods result in more precise estimates that are less sensitive to numerical instability due to both strong correlations and small changes in the input data. We apply the methods to demonstrate the most likely causal risk factor for stroke at the chemokine gene cluster is monocyte chemoattractant protein-1.

KEYWORDS

causal inference, correlated variants, dimension reduction, gene cluster, Mendelian randomization

1 | INTRODUCTION

Genetic variants associated with molecular and phenotypic traits can provide evidence on the causal pathways linking the associated trait with a disease outcome (Burgess et al., 2018). Various analytical approaches, including Mendelian randomization and colocalization, have been proposed that synthesize data on genetic associations to assess the nature of the relationship between a trait and a disease. In Mendelian randomization, it is assumed that the only pathway by which selected genetic variants influence the outcome is via the associated trait (Lawlor et al., 2008). Formally, genetic variants are assumed to satisfy the assumptions of an instrumental variable (Didelez & Sheehan, 2007). If multiple independent variants associated with the same trait show a consistent pattern of associations with the outcome, then it is plausible that the trait has a causal effect on the outcome (Bowden et al., 2016; Lawlor et al., 2016).

We here consider an extension to standard Mendelian randomization known as multivariable Mendelian randomization, which allows genetic variants to be associated with multiple related traits (Burgess & Thompson, 2015). For instance, it is difficult to find specific genetic predictors of fat-free mass that are not also associated with fat mass. Multivariable Mendelian randomization can be implemented by fitting a multivariable regression model using genetic associations with each of the traits as predictors (Sanderson et al., 2019). Coefficients from this model represent direct effects; that is, the effect of varying one of the traits in the model while keeping other traits constant (Burgess, Thompson, et al., 2017; Carter et al., 2021). Such investigations have suggested that fat mass rather than fat-free mass influences cardiovascular disease risk (Larsson et al., 2020), and that, amongst a set of lipid traits, apolipoprotein B is the primary driver of coronary heart disease risk (Zuber et al., 2021).

Often, Mendelian randomization investigations are conducted using genetic variants from a single genetic region, an approach known as cis-Mendelian randomization (Schmidt et al., 2020). This approach is particularly common when the risk factor is a gene product, such as gene expression or circulating levels of a protein. Such analyses are somewhat fragile, as the evidence is based on a single genetic region and so it is not possible to assess heterogeneity of findings across multiple genetic regions that represent independent datapoints (Burgess et al., 2020). However, if the function of the gene is well-understood, these analyses can be particularly insightful into the impact of intervening on a specific biological pathway. In some cases, the function of the gene may correspond to the action of a pharmacological agent, and hence the analysis is informative about a druggable pathway (Gill et al., 2021). Examples include the use of variants in the *HMGCR* gene region to inform about the impact of statin drugs (Ference et al., 2015), and variants in the *IL6R* gene region about the impact of interleukin-6 receptor inhibitors, such as tocilizumab (IL6R Genetics Consortium & Emerging Risk Factors Collaboration, 2012).

However, some genetic regions contain multiple genes (referred to as a gene cluster), and so are associated with multiple gene products. For example, the *FADS* locus contains genetic predictors of various fatty acids (Lattka et al., 2010), and the *IL1RL1–IL18R1* locus (the interleukin-1 receptor cluster) contains protein quantitative trait loci (pQTLs) for several proteins (Sun et al., 2018). Although variants in the interleukin-1 cluster are associated with several autoimmune diseases (Timms et al., 2004; G. Zhu et al., 2008), it is difficult to determine which of the proteins are causal risk factors (Reijmerink et al., 2010). Although a multivariable cis-Mendelian randomization approach has been proposed to disentangle complex gene regions and identify the causal drivers of disease (Porcu et al., 2019), authors of this approach suggest pruning genetic variants to near independence ($r^2 < 0.1$) to avoid potential problems of collinearity. However, it may not be possible to find sufficient near-independent variants for the multivariable regression model to give precise estimates for each trait. Although it is possible to prune at a less strict threshold, we have previously shown that under-pruning can result in ill-conditioning (Burgess, Zuber, et al., 2017). This represents a Goldilocks dilemma: too much pruning and we get imprecision or even lack of identification; too little pruning and we can get results that are highly sensitive to small changes in the estimated correlation matrix, and can be nonsensical.

We propose two methods for multivariable cis-Mendelian randomization that perform dimension reduction on the available genetic variants at a locus using principal component analysis (PCA). These methods reduce information on large numbers of highly correlated variants into a smaller number of orthogonal components, allowing efficient multivariable analyses to be implemented that are not so sensitive to high correlations between variants or small changes in the inputs. We demonstrate the superior performance of these methods over pruning methods in a simulation study, and illustrate the methods in a applied analysis investigating effects on stroke risk of three cytokines associated with a gene cluster on chromosome 17.

2 | METHODS

2.1 | Overview of the approach

Multivariable Mendelian randomization takes genetic variants that are each associated with at least one of a set of related exposure traits, and satisfy the instrumental variable assumptions for multivariable Mendelian randomization:

i. each variant is associated with one or more of the exposures,
ii. each exposure is associated with one or more of the genetic variants,
iii. each variant is not associated with the outcome via a confounding pathway, and
iv. each variant does not affect the outcome directly, only possibly indirectly via one or more of the exposures.

Although the approach was originally developed for use with individual-level data using the established two-stage least squares method (Burgess & Thompson, 2015), equivalent estimates can be obtained using summarized genetic association data that are typically reported by genome-wide association studies (GWAS) (Sanderson et al., 2019). We use summarized genetic association data (Bowden et al., 2017), and denote the genetic association of variant j with exposure trait k as $\hat{\beta}_{Xjk}$; this is the beta-coefficient from univariable regression of the trait on the variant. We denote the genetic association of variant j with the outcome as $\hat{\beta}_{Yj}$ and its standard error as $se(\hat{\beta}_{Yj})$; again, this is obtained from regression of the outcome on the variant.

2.2 | Inverse-variance weighted method

If the genetic variants are uncorrelated, then multivariable Mendelian randomization estimates can be obtained by fitting a multivariable model using weighted linear regression:

$$\begin{aligned} \hat{\beta}_{Yj} &= \theta_1 \hat{\beta}_{Xj1} + \theta_2 \hat{\beta}_{Xj2} + \ldots + \theta_K \hat{\beta}_{XjK} + \epsilon_j \\ \epsilon_j &\sim \mathcal{N}(0, (\hat{\beta}_{Yj})^2) \end{aligned} \tag{1}$$

for variants $j = 1, 2, \ldots, J$, where K is the total number of traits ($K > J$), and the error terms ϵ_j have independent normal distributions (Burgess et al., 2015). The parameter θ_k is an estimate of the direct effect of the kth trait on the outcome (i.e., the effect of intervening on that trait while keeping all other traits unchanged) (Carter et al., 2021). We refer to this method as the multivariable inverse-variance weighted (MV-IVW) method, as it is an extension of the univariable IVW method (Burgess et al., 2013) to the multivariable setting.

If the genetic variants are correlated, then we allow the error terms to be correlated and use generalized weighted linear regression:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_Y &= \theta_1 \hat{\boldsymbol{\beta}}_{X1} + \theta_2 \hat{\boldsymbol{\beta}}_{X2} + \ldots + \theta_K \hat{\boldsymbol{\beta}}_{XK} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \Sigma), \end{aligned} \tag{2}$$

where bold face represents vectors, and Σ is a variance-covariance matrix with elements $\Sigma_{j_1,j_2} = \text{se}(\hat{\beta}_{Yj_1})\, \text{se}(\hat{\beta}_{Yj_2})\, \rho_{j_1,j_2}$, with ρ_{j_1,j_2} representing the correlation between the j_1th and j_2th variants. This method was advocated by Porcu et al. (2019) for the analysis of summarized genetic association data on gene expression traits with shared genetic predictors. Estimates are obtained as

$$\hat{\theta}_{\mathbf{MV}} - \mathbf{IVW} = (\hat{\boldsymbol{\beta}}_X{}^T \Sigma^{-1} \hat{\boldsymbol{\beta}}_X)^{-1} \hat{\boldsymbol{\beta}}_X{}^T \Sigma^{-1} \hat{\boldsymbol{\beta}}_Y, \tag{3}$$

where $\hat{\boldsymbol{\beta}}_X$ is the J by K matrix of genetic associations with the traits, and $\hat{\boldsymbol{\beta}}_Y$ is the J by 1 vector of genetic associations with the outcome. We note that this method is identical to the method referred to as transcriptome-wide Mendelian randomization (TWMR) by Porcu et al. (2019).

This calculation requires inversion of the variance-covariance matrix Σ, which can lead to numerical instability if the matrix of correlations between genetic variants is near-singular. This occurs when there is a set of genetic variants that is close to being linearly dependent. If a set of genetic variants is linearly dependent (i.e., there is at least one variant that can be perfectly predicted based on the other variants), then the correlation matrix will be exactly singular, and so cannot be inverted. If a set of genetic variants is almost but not exactly linearly dependent, then the correlation matrix can be inverted, but some elements of the matrix inverse will be very large in magnitude. This results in an ill-conditioned problem, meaning that small changes in the inputs can lead to large changes in the estimates. The condition number of a matrix is a measure of ill-conditioning; for a positive-definite symmetric matrix, this can be calculated as the ratio of the largest to the smallest eigenvalue. Although there are no universal thresholds, condition numbers over 100 are cause for concern, particularly if the genetic associations are known to a limited degree of precision.

To implement the proposed PCA method, we first consider the matrix Ψ where:

$$\Psi_{j_1,j_2} = \sum_k \left| \hat{\beta}_{Xj_1k} \right| \sum_k \left| \hat{\beta}_{Xj_2k} \right| \text{ se}\left(\hat{\beta}_{Yj_1}\right)^{-1} \text{se}\left(\hat{\beta}_{Yj_2}\right)^{-1} \rho_{j_1,j_2}. \tag{4}$$

This is a weighted version of the variance-covariance matrix, with weights taken as the sum of the absolute values of the genetic associations with the traits. Obtaining principal components of this matrix ensures that the top principal components assign greater weights for variants having larger associations with the traits and more precise associations with the outcome. If genetic associations with the outcome are all estimated in the same data set, then their standard errors will depend on the minor allele frequency of the variants; inclusion of these standard errors in the weighting matrix prioritizes genetic variants that have more precise associations with the outcome, and hence provide the most information to the Mendelian randomization estimate. Considering the PCA decomposition $\Psi = W \Lambda W^T$, where W is the matrix of eigenvectors (or loadings) and Λ is the diagonal matrix with the eigenvalues $\lambda_1 > \ldots > \lambda_J$ on the diagonal, let W_k be the matrix constructed of the first k columns of W. Then we define:

$\tilde{\boldsymbol{\beta}}_X = W_k^T \hat{\boldsymbol{\beta}}_X$ as the matrix of transformed genetic associations with the exposure traits

$\tilde{\boldsymbol{\beta}}_Y = W_k^T \hat{\boldsymbol{\beta}}_Y$ as the vector of transformed genetic associations with the outcome

$\tilde{\Sigma} = W_k^T \Sigma W_k$ as the transformed variance-covariance matrix.

The multivariable inverse-variance weighted principal component analysis (MV-IVW-PCA) estimate is given by

$$\hat{\theta}_{MV-IVW-PCA} = (\tilde{\boldsymbol{\beta}}_X{}^T \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\beta}}_X)^{-1} \tilde{\boldsymbol{\beta}}_X{}^T \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\beta}}_Y. \tag{5}$$

This is an adaptation of the MV-IVW method using transformed genetic instruments that represent linear weighted scores comprised of the original genetic variants, where the weights of the scores are the eigenvectors from the PCA decomposition. As the principal components are orthogonal, the transformed variance-covariance matrix should not be near-singular. The standard errors of these estimates are:

$$\mathrm{se}(\hat{\theta}_{k,MV-IVW-PCA}) = \sqrt{(\tilde{\boldsymbol{\beta}}_X{}^T \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\beta}}_X)^{-1}_{kk}}. \tag{6}$$

In our investigations, we set the number of principal components to explain 99% of the variance in the matrix Ψ.

2.3 | Limited information maximum likelihood method (LIML)

An alternative method for instrumental variable analysis is the LIML method (Baum et al., 2007). The estimate from the LIML method minimizes the Anderson–Rubin statistic (Anderson & Rubin, 1949), which is a measure of heterogeneity of the estimates based on the different genetic variants. Both the IVW and LIML methods are part of a larger family of methods, known as the generalized method of moments (GMM) (Hansen, 1982). We here derive a multivariable analogue of the LIML method that can be implemented using summarized genetic association data for correlated variants. We refer to this method as the multivariable limited information maximum likelihood (MV-LIML) method.

For the MV-LIML method, we require the additional knowledge of the $K \times K$ correlation matrix of exposures, which we denote Φ. In the simulation study, we set this matrix to be the identity.

We let $\mathbf{g}(\theta) = \hat{\boldsymbol{\beta}}_Y - \hat{\boldsymbol{\beta}}_X\, \theta$, where $\theta = (\theta_1, \ldots, \theta_K)^T$. Under the assumption that all J variants are valid instruments, setting $\mathbf{g}(\theta) = \mathbf{0}$ provides a set of J estimating equations for the K unknown parameters θ_k. When $J > K$, if the genetic variants are linearly independent, it is generally not possible to find an estimator $\hat{\theta}$ that can set $\mathbf{g}(\hat{\theta}) = \mathbf{0}$. as the number of equations is greater than the number of unknown parameters. Thus, LIML-based methods take K linear combinations of the J estimating equations, where the weights in the linear combination are chosen to minimize the variance of the resulting estimator $\hat{\theta}$.

In particular, the MV-LIML estimator is given by

$$\hat{\theta}_{MV-LIML} = \arg\min_{\theta} \hat{Q}(\theta), \tag{7}$$

where $Q(\theta) = \mathbf{g}(\theta)^T\, \Omega(\theta)^{-1}\, \mathbf{g}(\theta)$, and $\Omega(\theta)$ is a $J \times J$ matrix with its (j_1, j_2)th element given by

$$\begin{aligned} \Omega_{j_1,j_2}(\theta) = {} & \left(\mathrm{se}\left(\hat{\beta}_{Y_{j_1}}\right)\mathrm{se}\left(\hat{\beta}_{Y_{j_2}}\right)\rho_{j_1,j_2}\right) \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sqrt{\mathrm{se}\left(\hat{\beta}_{X_{j_1 k}}\right)\mathrm{se}\left(\hat{\beta}_{X_{j_2 k}}\right)} \\ & \sqrt{\mathrm{se}\left(\hat{\beta}_{X_{j_1 l}}\right)\mathrm{se}\left(\hat{\beta}_{X_{j_2 l}}\right)}\,\rho_{j_1,j_2}\Phi_{k,l}\theta_k\theta_l. \end{aligned} \tag{8}$$

For exposure k, the MV-LIML estimator of θ_k is given by the kth element of $\hat{\theta}_{MV-LIML}$, and its standard error is given by $\sqrt{\hat{V}_{k,k}}$, where $\hat{V}_{k,k}$ is the kth diagonal element of the $K \times K$ matrix $\hat{V} = (\hat{\boldsymbol{\beta}}_X{}^T\; \Omega(\hat{\theta}_{MV-LIML})^{-1}\; \hat{\boldsymbol{\beta}}_X)^{-1}$.

Theoretical results suggest LIML provides robust estimation when using many weak instruments (Chao & Swanson, 2005). For univariable cis-Mendelian randomization analyses, simulation evidence has further highlighted the low bias properties of LIML-based estimators in finite samples (Patel et al., 2020).

We consider a version of the LIML method using PCA to perform dimension reduction on the set of genetic variants by replacing:

$$\begin{aligned}\tilde{\mathbf{g}}(\theta) &= W_k^T\,\mathbf{g}(\theta)\\ \tilde{\Omega}(\theta) &= W_k^T\,\Omega(\theta)\,W_k\\ \tilde{Q}(\theta) &= \tilde{\mathbf{g}}(\theta)^T\;\tilde{\Omega}(\theta)^{-1}\;\tilde{\mathbf{g}}(\theta)\end{aligned} \tag{9}$$

and minimizing $\tilde{Q}(\theta)$, with W_k defined as previously. We refer to this method as MV-LIML-PCA. As per the MV-IVW-PCA method, we set the number of principal components to explain 99% of the variance in the matrix Ψ.

2.4 | Simulation study

We perform a simulation study to assess whether the proposed PCA methods are able to detect which out of a set of traits with shared clustered genetic predictors influences an outcome, and whether the causal effects of the traits can be estimated reliably.

We consider a scenario with three traits and 100 correlated genetic variants. We generate 10 000 simulated datasets according to the following data-generating model for 20 000 individuals indexed by i:

$$\begin{aligned}A_{j_1,j_2} &\sim \text{Uniform}(-0.3, 1) \text{ for } j_1, j_2 = 1, \ldots, 100\\ B &= \text{cor}(A\;A^T)\\ \mathbf{G_i} &\sim \mathcal{N}_J(\mathbf{0}, B)\\ \alpha_j &\sim \mathcal{N}(0.08, 0.01^2) \text{ for } j = 1, \ldots, 15\\ X_{i1} &= \sum_{j=1}^{5} \alpha_j G_{ij} + U_{i1} + \epsilon_{Xi1}\\ X_{i2} &= \sum_{j=6}^{10} \alpha_j G_{ij} + U_{i2} + \epsilon_{Xi2}\\ X_{i3} &= \sum_{j=11}^{15} \alpha_j G_{ij} + U_{i3} + \epsilon_{Xi3}\\ Y_i &= 0.4X_{i1} - 0.6X_{i3} + U_{i1} + U_{i2} + U_{i3} + \epsilon_{Yi}\\ U_{i1}, U_{i2}, U_{i3}, \epsilon_{Xi1}, \epsilon_{Xi2}, \epsilon_{Xi3}, \epsilon_{Yi} &\sim \mathcal{N}(0, 1)\text{independently.}\end{aligned}$$

The genetic variants G_j are simulated from a multivariable normal distribution with mean zero and variance-covariance matrix B. The traits X_1, X_2, and X_3 are simulated such that 5 variants influence the first trait, the next 5 influence the second trait, the next 5 influence the third trait, and the remaining 85 do not influence any trait. However, due to the moderately large correlations between the genetic variants (which typically range from around −0.1 to +0.6 with an interquartile range from around +0.2 to +0.4), typically each of the 100 variants is associated with all three traits at a genome-wide level of significance. The outcome Y is influenced by traits X_1 and X_3, with the true effect of X_1 set at +0.4 and the true effect of X_3 set at −0.6. The associations between the traits and the outcome are affected by confounders U_1, U_2, and U_3.

We estimate genetic associations with the exposures on the first 10,000 individuals, and genetic associations with the outcome on the subsequent 10,000 individuals. Correlations between the genetic variants are estimated on the first 10,000 individuals. This represents a two-sample scenario, where genetic associations with the exposures and outcome are obtained on nonoverlapping samples (Pierce & Burgess, 2013). The mean instrument strength based on the 5 causal variants for each trait is $R^2 = 3.5\%$, corresponding to a mean univariable F statistic (on 5 and 19,994 degrees of freedom) for each trait of around 145, and a conditional F statistic (on 15 and 19,984 degrees of freedom) for each trait of around 22 (Sanderson et al., 2019). Although these F statistics are the most relevant measure of instrument strength for the oracle analyses, instrument strength in all other analyses will depend on the number and identity of the genetic variants included in each analysis, which will vary between simulations and depending on the pruning threshold.

We compare four different methods: the MV-IVW and MV-LIML methods with various choices of genetic variants as inputs, and the MV-IVW-PCA and MV-LIML-PCA methods described above. For the MV-IVW and MV-LIML methods, we consider pruning the variants at thresholds of $|\rho| < 0.4$, $|\rho| < 0.6$, and $|\rho| < 0.8$ (equivalent to $r^2 < 0.16$, $r^2 < 0.36$, and $r^2 < 0.64$). We note that pruning at $|\rho| < 0.1$ would often result in 3 or fewer variants being available for analysis, which would not allow multivariable Mendelian randomization to be attempted, as the number of genetic variants needs to be greater than the number of traits. Pruning is performed by first selecting the variant with the lowest p value for association with any of the traits, and then excluding from consideration all variants more strongly correlated with the selected variant than the threshold value. We then select the variant amongst those remaining with the lowest p value, and exclude variants more correlated with that variant than the threshold value. We repeat until all variants have either been selected or excluded. We also consider an oracle choice of variants, in which only the 15 genetic variants that truly influence the traits are used as instruments.

In addition to the main simulation study, we also consider the performance of methods with other parameter settings: (1) weaker instruments: we generate the α_j parameters from a normal distribution with mean 0.05 (corresponding to a mean instrument strength of $R^2 = 1.4\%$, mean univariable $F = 57$, mean conditional $F = 9.4$); (2) stronger correlations: we generate elements of the A matrix from a uniform distribution on +0.1 to +1.0, resulting in correlations which typically range from around +0.5 to +0.85 with an interquartile range from around +0.65 to +0.75; (3) stronger causal effects, with $\theta_1 = +0.8$ and $\theta_3 = +1.0$, and (4) two alternative approaches for generating a correlation matrix taken from the R package *clusterGeneration* (Joe, 2006), (4) real linkage disequilibrium data: a correlation matrix estimated in UK Biobank participants of European ancestries on variants from the chemokine receptor gene cluster, pruned at a threshold of $r^2 < 0.95$ to avoid very high correlations; and (5) correlated exposures: the exposures were correlated by additionally adding U_{i2} to X_1, U_{i3} to X_2, and U_{i1} to X_3. Further details of these additional scenarios are provided in the Supporting Information.

In the main simulation study, we also consider a conditional selection of genetic variants, in which we first select the variant having the lowest p value for association with any of the traits. In each subsequent step, we select the variant having the lowest p value for association with any of the traits conditional on all previously selected variants. We repeat until no additional variant is conditionally associated with any trait below a threshold of $p < 0.001$. Although this approach does not guarantee that variants will not be highly correlated, it is unlikely that two very highly correlated variants would both be conditionally associated with a trait. For convenience, in the main simulation study, we implement this method using individual-level data. If we only had access to summarized data, a mathematically equivalent procedure could be implemented using the Genome-wide Complex Trait Analysis conditional & joint association analysis (GCTA-COJO) method (Yang et al., 2012).

Further, we consider two variations to the main simulation study that are reflective of potential problems that may arise in applied practice. First, we estimate the variant correlation matrix based on an independent sample. This reflects the common occurrence that the correlation matrix is obtained from a reference sample rather than the data set under analysis, and assesses robustness of the methods to variability in the correlation matrix. Second, we round the genetic associations and their standard errors to three decimal places. This reflects the common occurrence that genetic associations are obtained from a publicly-available source, and hence are not known to absolute precision. Again, this assesses robustness of the methods to variability in the data inputs.

2.5 | Applied example: Chemokine gene cluster and risk of stroke

We illustrate our methods using data on genetic associations with three cytokines and stroke risk. Previous research has implicated monocyte chemoattractant protein-1 (MCP-1), which is also called chemokine (C-C motif) ligand 2 (CCL2), in the pathophysiology of stroke (Georgakis, De Lemos, et al., 2021; Georgakis, Gill, et al., 2019; Georgakis, van der Laan, et al., 2021). However, the *CCL2* gene that encodes this cytokine is located in a cluster that also includes genes *CCL7*, and *CCL11*. Variants in this genetic region are associated with multiple cytokines other than MCP-1, including MCP-3 (also called CCL7), and eotaxin-1 (also called CCL11). Hence, it is not clear from univariable Mendelian randomization (i.e., analyses with a single exposure trait) which of these proteins is driving stroke risk.

We conduct a multivariable cis-Mendelian randomization analysis to disentangle the effects of these cytokines. We take variants from the *CCL2* gene region (GRCh38/hg38; chr17:34,255,218–34,257,203) plus 500 kilobasepairs either side, genetic associations with the cytokines from a reanalysis of data on three Finnish cohorts by Ahola-Olli et al. (2017) that did not adjust for body mass index (Kalaoja et al., 2021), and genetic associations with all stroke and cardioembolic stroke from European ancestry individuals in the MEGASTROKE consortium (Malik et al., 2018). Cardioembolic stroke was chosen as genetic associations were stronger with this stroke subtype than for all stroke in a motivating Mendelian randomization analysis that included variants from throughout the genome (Georgakis, Gill, et al., 2019). Correlations between variants were estimated in 376 703 individuals of European ancestries from UK Biobank.

3 | RESULTS

3.1 | Simulation study

Results from the simulation study are shown in Table 1. For each method, we display the mean estimate of each parameter, the standard deviation of estimates, the mean standard error, and the empirical power of the 95% confidence interval, which represents the proportion of confidence intervals that exclude the null. For θ_2, the empirical power is the Type 1 error rate, and should be close to 5%.

Although the MV-IVW and MV-LIML methods perform well under the oracle setting, power to detect a causal effect is substantially reduced when pruning variants at 0.4. Although increasing the pruning threshold to 0.6 increases power, it also results in Type 1 error inflation. For the MV-IVW method, Type 1 error rates increase to 9.1%, and for the MV-LIML method, to 20.2%. When increasing the pruning threshold to 0.8, estimates are completely unreliable, with mean estimates in the MV-IVW method for θ_1 and θ_3 having the wrong sign.

In contrast, the MV-IVW-PCA and MV-LIML-PCA methods perform well throughout, with Type 1 error rates similar to those of the oracle methods, and greater power to detect a causal effect than the methods that rely on pruning. For the MV-IVW-PCA method, the variability and mean standard errors of estimates are similar to those of the oracle methods, although estimates of θ_1 and θ_3 are attenuated. This is a result of weak instrument bias, which affects analyses similarly to non-differential measurement error. With a single exposure in a two-sample setting, estimates are biased towards the null due to imprecision in the estimated genetic associations with the exposures. With multiple exposures, this bias may act in any direction, even in a two-sample setting (J. Zhu et al., 2022).

Similar findings were observed when considering weaker instruments (Supporting Information: Table A2), stronger correlations (Supporting Information: Table A3), stronger causal effects (Supporting Information: Table A2), alternative synthetic correlation matrices (Supporting Information: Tables A4 and A5), a correlation matrix based on real genetic data (Supporting Information: Table A6), and with correlated exposures (Supporting Information: Tables A7–A9). Although the power varied between simulation settings, in each case the PCA methods outperformed the pruning methods in terms of power and precision at a threshold of 0.4, whereas at a threshold of 0.6 the MV-IVW and MV-LIML methods had inflated Type 1 error rates. Type 1 error rates for the PCA methods were generally well controlled, although the Type 1 error for the MV-LIML-PCA method was slightly inflated with weaker instruments (6.7% for MV-IVW-PCA, 13.9% for MV-LIML-PCA), with correlated exposures (moderate correlations: 7.5% for MV-IVW-PCA, 10.3% for MV-LIML-PCA; weaker correlations: 7.6% for MV-IVW-PCA, 10.9% for MV-LIML-PCA; stronger correlations: 7.5% for MV-IVW-PCA, 10.0% for MV-LIML-PCA), and the Type 1 error for the MV-IVW-PCA method was slightly inflated with stronger causal effects (12.8% for MV-IVW-PCA, 8.2% for MV-LIML-PCA). This highlights the importance of specifying these trait correlations in the MV-LIML-PCA method, which were set at zero in the simulation study for simplicity. For the correlated exposure scenarios, we repeated the MV-LIML-PCA method accounting for correlations between the exposures: Type 1 error rates were reduced to 9.1% (moderate correlations), 10.3% (weaker correlations), and 8.2% (stronger correlations).

We also considered a conditional approach for the selection of genetic variants. As this approach requires conditional associations with each exposure for each variant to be recalculated at each step of the variant selection algorithm, this approach is considerably more computationally intensive than the pruning approach, and so we only considered estimates for the first 1000 simulated datasets in the main simulation. Results are provided in Table 2. The conditional selection method performed better than the pruning methods, although it was outperformed by the MV-IVW-PCA method in terms of variability and precision of estimates (the conditional selection method had greater standard deviations and mean standard errors), and by the MV-LIML-PCA method in terms of power to detect a causal effect.

We also considered two additional variations to the main simulation reflective of potential problems in applied practice. Table 3 shows results in which the variant correlation matrix was obtained from an independent sample of 10,000 individuals. Results were similar, except that the Type 1 error rate for the MV-IVW method at a pruning threshold of 0.6 was slightly higher at 11.2%. When obtaining the correlation matrix from an independent sample of 1000 individuals, Type 1 error rates for the MV-IVW method were higher still at 18.7% (Supporting Information: Table A10). Table 4 shows results in which the genetic association estimates were rounded to 3 decimal places. Again, results were similar, except that the Type 1 error rate for the MV-IVW method at a pruning threshold of 0.6 was notably higher at 15.1%. In contrast, results from the PCA methods were not sensitive to changes in the variant correlation matrix or rounding of the genetic association estimates.

3.2 | Applied example: Chemokine gene cluster and risk of stroke

Genetic associations with each of the cytokines and all stroke were available for 2922 variants, and with cardioembolic stroke for 2904 variants. We compare results from the MV-IVW-PCA method to those from the MV-IVW method at a pruning threshold of $|\rho| < 0.1$ (equivalent to $r^2 < 0.01$), $|\rho| < 0.4$ (equivalent to $r^2 < 0.16$), and $|\rho| < 0.6$ (equivalent to $r^2 < 0.13$). In the MV-IVW-PCA method, we initially pruned at $|\rho| < 0.95$ to remove very highly correlated variants from the analysis. We also excluded variants not associated with any of the cytokines at $p < 0.001$ from all analyses. For the pruned set of variants at $|\rho| < 0.95$, the genetic

TABLE 1 Results from the main simulation study

Parameter	Method	Pruning	Mean	SD	Mean SE	Power
θ_1	MV-IVW	Oracle	0.353	0.133	0.120	81.4
		0.4	0.304	0.164	0.147	57.4
		0.6	0.207	0.115	0.094	60.9
		0.8	−0.083	0.417	0.051	76.5
	MV-LIML	Oracle	0.379	0.143	0.133	81.4
		0.4	0.340	0.188	0.163	58.9
		0.6	0.316	0.212	0.103	77.0
		0.8	0.083	2.372	0.179	78.8
	MV-IVW-PCA	–	0.296	0.130	0.119	69.3
	MV-LIML-PCA	–	0.347	0.152	0.130	74.5
θ_2	MV-IVW	Oracle	−0.005	0.133	0.120	7.4
		0.4	−0.012	0.166	0.147	7.5
		0.6	−0.003	0.112	0.094	9.1
		0.8	0.037	0.408	0.051	75.4
	MV-LIML	Oracle	−0.001	0.144	0.133	6.2
		0.4	−0.007	0.192	0.163	7.3
		0.6	0.006	0.186	0.103	20.2
		0.8	0.010	2.522	0.179	76.6
	MV-IVW-PCA	–	−0.013	0.129	0.119	7.1
	MV-LIML-PCA	–	−0.006	0.154	0.130	8.7
θ_3	MV-IVW	Oracle	−0.545	0.132	0.120	98.6
		0.4	−0.487	0.166	0.148	87.6
		0.6	−0.315	0.139	0.094	86.9
		0.8	0.220	0.418	0.051	77.8
	MV-LIML	Oracle	−0.576	0.142	0.134	98.8
		0.4	−0.531	0.190	0.164	88.6
		0.6	−0.451	0.212	0.103	92.6
		0.8	0.005	2.250	0.179	79.3
	MV-IVW-PCA	–	−0.476	0.130	0.119	96.1
	MV-LIML-PCA	–	−0.538	0.152	0.131	97.0

Note: Mean estimates, standard deviation (SD) of estimates, mean standard error (mean SE) of estimates, and empirical power of the 95% confidence interval to estimate $\theta_1 = 0.4$, $\theta_2 = 0$, and $\theta_3 = -0.6$. We consider four methods, and various pruning thresholds for the MV-IVW and MV-LIML methods, plus an oracle setting in which only the 15 variants that truly affect the traits are included in the analysis.

associations with the different cytokines were not highly correlated: correlations were 0.21 between genetic associations with MCP-1 and MCP-3; 0.27 between genetic associations with MCP-1 and eotaxin-1; and 0.01 between genetic associations with MCP-3 and eotaxin-1. Estimates for the three cytokines, which represent log odds ratios per 1 standard deviation increase in the cytokine, are provided in Table 5.

For all stroke, at a pruning threshold of 0.1, the MV-IVW method indicates that MCP-1 is the true causal risk factor. However, a researcher may be tempted to consider a less strict pruning threshold to obtain more precise estimates. But at a pruning threshold of 0.4, the MV-IVW method indicates that eotaxin-1 is the true causal risk factor, and at a pruning threshold of 0.6, the MV-IVW method again indicates that MCP-1 has the strongest

TABLE 2 Results from the main simulation study for conditional selection method

Parameter	Method	Mean	SD	Mean SE	Power
θ_1	Oracle	0.357	0.129	0.120	82.5
	Cond select	0.334	0.137	0.126	73.3
	MV-IVW-PCA	0.299	0.126	0.118	70.8
	MV-LIML-PCA	0.349	0.150	0.130	76.3
θ_2	Oracle	−0.012	0.133	0.121	7.2
	Cond select	−0.015	0.140	0.126	7.4
	MV-IVW-PCA	−0.018	0.126	0.119	5.8
	MV-LIML-PCA	−0.009	0.148	0.131	7.1
θ_3	Oracle	−0.543	0.130	0.121	98.7
	Cond select	−0.515	0.142	0.127	96.5
	MV-IVW-PCA	−0.475	0.130	0.119	95.9
	MV-LIML-PCA	−0.539	0.151	0.131	96.6

Note: Results from oracle, conditional selection, MV-IVW-PCA, and MV-LIML-PCA methods across 1000 simulated datasets: mean estimates, standard deviation (SD) of estimates, mean standard error (mean SE) of estimates, and empirical power of the 95% confidence interval to estimate $\theta_1 = 0.4$, $\theta_2 = 0$, and $\theta_3 = -0.6$.

TABLE 3 Results from the main simulation study with a correlation matrix estimated in an independent sample of 10,000 individuals

Parameter	Method	Pruning	Mean	SD	Mean SE	Power
θ_1	MV-IVW	0.4	0.303	0.166	0.147	57.1
		0.6	0.206	0.114	0.090	62.4
	MV-LIML	0.4	0.340	0.189	0.162	59.0
		0.6	0.301	0.183	0.098	76.7
	MV-IVW-PCA	–	0.295	0.130	0.118	69.2
	MV-LIML-PCA	–	0.347	0.153	0.130	74.7
θ_2	MV-IVW	0.4	−0.013	0.167	0.148	7.3
		0.6	−0.005	0.114	0.090	11.2
	MV-LIML	0.4	−0.009	0.193	0.163	7.3
		0.6	0.003	0.174	0.098	20.8
	MV-IVW-PCA	–	−0.012	0.129	0.118	7.2
	MV-LIML-PCA	–	−0.006	0.154	0.130	8.9
θ_3	MV-IVW	0.4	−0.485	0.165	0.148	86.8
		0.6	−0.322	0.126	0.090	88.9
	MV-LIML	0.4	−0.528	0.188	0.164	87.9
		0.6	−0.436	0.207	0.099	93.6
	MV-IVW-PCA	–	−0.476	0.130	0.119	96.1
	MV-LIML-PCA	–	−0.538	0.152	0.131	97.0

Note: Mean estimates, standard deviation (SD) of estimates, mean standard error (mean SE) of estimates, and empirical power of the 95% confidence interval to estimate $\theta_1 = 0.4$, $\theta_2 = 0$, and $\theta_3 = -0.6$.

TABLE 4 Results from the main simulation study with genetic associations ($\boldsymbol{\beta}$-coefficients and standard errors) rounded to three decimal places

Parameter	Method	Pruning	Mean	SD	Mean SE	Power
θ_1	MV-IVW	0.4	0.305	0.165	0.147	57.6
		0.6	0.200	0.138	0.090	59.7
	MV-LIML	0.4	0.341	0.187	0.162	59.4
		0.6	0.302	0.192	0.099	75.6
	MV-IVW-PCA	–	0.296	0.130	0.119	69.2
	MV-LIML-PCA	–	0.346	0.153	0.130	74.7
θ_2	MV-IVW	0.4	−0.013	0.165	0.148	7.3
		0.6	−0.012	0.131	0.090	15.1
	MV-LIML	0.4	−0.009	0.190	0.163	7.1
		0.6	−0.001	0.183	0.099	24.4
	MV-IVW-PCA	–	−0.013	0.129	0.119	7.1
	MV-LIML-PCA	–	−0.006	0.154	0.130	8.8
θ_3	MV-IVW	0.4	−0.487	0.166	0.148	87.5
		0.6	−0.330	0.145	0.090	89.3
	MV-LIML	0.4	−0.529	0.189	0.164	88.4
		0.6	−0.459	0.195	0.099	94.3
	MV-IVW-PCA	–	−0.476	0.131	0.119	96.0
	MV-LIML-PCA	–	−0.536	0.152	0.131	96.9

Note: Mean estimates, standard deviation (SD) of estimates, mean standard error (mean SE) of estimates, and empirical power of the 95% confidence interval to estimate $\theta_1 = 0.4$, $\theta_2 = 0$, and $\theta_3 = -0.6$.

TABLE 5 Applied example: effect of three cytokines on stroke risk

				MCP-1		MCP-3		Eotaxin-1	
Method	**Pruning**	**Variants/ PCs**	**Cond number**	**Estimate (SE)**	**p Value**	**Estimate (SE)**	***p* Value**	**Estimate (SE)**	***p* Value**
All stroke									
MV-IVW	0.1	20	27.7	0.091 (0.045)	0.046	−0.062 (0.046)	0.18	0.062 (0.082)	0.45
	0.4	75	1224	0.057 (0.035)	0.11	−0.014 (0.024)	0.55	0.110 (0.050)	0.028
	0.6	151	17762	−0.038 (0.022)	0.09	−0.014 (0.017)	0.41	0.040 (0.029)	0.17
MV-IVW-PCA		30	24.9	0.075 (0.041)	0.071	−0.029 (0.027)	0.29	0.000 (0.063)	0.99
Cardioembolic stroke									
MV-IVW	0.1	19	22.7	0.270 (0.095)	0.005	0.151 (0.104)	0.14	−0.174 (0.169)	0.30
	0.4	70	870	0.141 (0.073)	0.053	0.003 (0.051)	0.96	−0.132 (0.107)	0.21
	0.6	145	15,790	0.089 (0.046)	0.056	0.040 (0.036)	0.27	0.019 (0.062)	0.76
MV-IVW-PCA		29	25.9	0.254 (0.089)	0.004	−0.018 (0.065)	0.78	−0.108 (0.145)	0.46

Note: Estimates (standard errors, SE) and *p* values from MV-IVW and MV-IVW-PCA methods. Variants/PCs indicates the number of genetic variants (MV-IVW method) or principal components (PCs, MV-IVW-PCA method) included in the analysis. Cond number indicates the condition number of the variance-covariance matrix Σ; larger numbers signal worse problems due to ill-conditioning. Estimates represent log odds ratios per 1 standard deviation increase in the cytokine.

evidence of being the true causal risk factor, but the causal estimate is in the other direction. In contrast, the MV-IVW-PCA method indicates that MCP-1 has the strongest evidence of being the true causal risk factor, similarly to the MV-IVW method at the most conservative pruning threshold. Compared with results at this threshold, estimates from the MV-IVW-PCA method have narrower standard errors, although the estimate for MCP-1 is slightly attenuated and so has a slightly higher p value. At a pruning threshold of 0.4, the condition number for the variance-covariance matrix Σ in the MV-IVW method is 1224, whereas the condition number for the transformed variance-covariance matrix $\tilde{\Sigma}$ in the MV-IVW-PCA method is 24.9; a larger number signals worse problems due to ill-conditioning. We would therefore trust results from the MV-IVW-PCA method and the MV-IVW method at a threshold of 0.1, which both suggest that the strongest evidence is for MCP-1 as the causal risk factor, and the effect is in the harmful direction. For cardioembolic stroke, estimates are more similar amongst the different implementations of the methods, with stronger evidence for MCP-1 as the causal risk factor at this locus, particularly from the MV-IVW-PCA method. These findings add to the existing body of basic science (Georgakis, van der Laan, et al., 2021), observational (Georgakis, De Lemos, et al., 2021; Georgakis, van der Laan, et al., 2019b), and genetic evidence (Georgakis, Gill, et al., 2019) implicating circulating MCP-1 in stroke risk.

4 | DISCUSSION

In this manuscript, we have introduced two methods for multivariable cis-Mendelian randomization, the MV-IVW-PCA and MV-LIML-PCA methods. Compared to existing methods that rely on pruning, these methods had superior performance: they outperformed pruning methods in terms of power to detect a causal effect, and they generally maintained close to nominal Type 1 error rates across a range of scenarios. They were also less sensitive that the pruning methods to variation in the variant correlation matrix, and to rounding of the genetic association estimates. We applied the MV-IVW-PCA method to disentangle the effects of three similar exposures with shared genetic predictors at a gene cluster; the method gave results that correspond to existing biological understanding of this pathway. We note that these methods are not designed for performing Mendelian randomization in most contexts, as most such analyses are polygenic, using genetic variants from across the genome. These methods are recommended for consideration when performing cis-Mendelian randomization, that is when genetic variants are taken from a single gene region; typically a gene region with functional relevance to the exposures of interest.

The approach of multivariable cis-Mendelian randomization has several potential applications. In our applied example, we considered proteins as exposures. Alternative potential applications could include expression of different genes as exposures, or expression of the same gene in different tissues. However, results from the latter case may be difficult to interpret if the genetic predictors of gene expression do not vary between tissues, or if data on variants affecting gene expression in all relevant tissues are not available. Another possible area of application is if there are different aspects of an exposure trait that could be considered as independent risk factors, such as concentration and isoform size of lipoprotein(a) (Saleheen et al., 2017). To obtain estimates for the different exposures, it is not necessary to have genetic predictors that are uniquely associated with each exposure trait, but it is necessary to have some variants that associate more or less strongly with the different traits (Sanderson et al., 2019).

An alternative approach for disentangling clusters of correlated variants associated with multiple traits is colocalization. Colocalization is a method that attempts to distinguish between two scenarios: a scenario in which two traits (which we here conceptualize as an exposure and an outcome) are influenced by distinct genetic variants, but there is overlap in the genetic associations due to correlation between the variants; and an alternative scenario in which the two traits are influenced by the same genetic variant (Wallace et al., 2012). In the latter scenario, it is likely that the two traits are causally related, although the colocalization method is agnostic as to whether the exposure trait influences the outcome trait, the outcome influences the exposure, or the two are influenced by a common causal factor (Solovieff et al., 2013). There are several conceptual differences between colocalization and cis-Mendelian randomization, although there are also similarities. Two specific advantages of the proposed cis-Mendelian randomization method are that it allows for the existence of multiple causal variants, in contrast to some colocalization methods (Foley et al., 2021; Giambartolomei et al., 2014), and it allows for the existence of multiple causal traits. Another feature is that it provides causal estimates, although the value of causal estimates in Mendelian randomization beyond indicating the direction of effect is disputed (VanderWeele et al., 2014).

Although the PCA methods can be implemented with highly correlated variants, we would recommend a minimal level of pruning (say, $r^2 < 0.9$) before applying the methods in practice as variants that are very highly correlated do not contribute independent information to the analysis, but could provide computational challenges. Even if two variants do not have a high pairwise correlation as measured by the r^2 statistic, they may be highly correlated. For example the case of "tagging variants," where a less common variant is only present when the common variant is present. This leads to collinearity even though the pairwise r^2 may be low. Another possibility is that three of more variants may perfectly or near-perfectly linear dependent even though their pairwise correlations are only moderate in size. This is particularly likely if the genetic variation in a region can be explained by a small number of haplotypes. Although the concept of ill-conditioning is not technically related to the problem of weak instruments (ill-conditioning occurs when genetic variants are highly correlated, weak instruments refers to the strength of genetic associations with the exposure), the phenomena are likely to be practically related, as if an investigator has strong instruments for a set of traits, then there is no strong motivation to include many genetic variants from a gene region in a cis-Mendelian randomization analysis. Whereas if the instruments are weak (or, in a multivariable context, conditionally weak [Sanderson & Windmeijer, 2016]), then an investigator is more likely to include multiple variants in an analysis in an attempt to bolster power.

Additionally, we have assumed in this paper that the traits under analysis are causally independent. If a trait has a causal effect on the outcome that is fully mediated by one of the other traits in the analysis, then the estimate for that trait would be zero. Hence, the method identifies the proximal causal risk factors for the outcome (Grant & Burgess, 2021). If there are large numbers of traits, the MV-IVW-PCA method could be combined with a Bayesian variable selection method that compares models with different sets of traits on the assumption of a sparse risk factor set (Zuber et al., 2020).

There are several limitations to these methods, which are shared by other methods for Mendelian randomization using summarized data (Bowden et al., 2017; Burgess et al., 2016). Uncertainty in genetic associations with the exposure traits is not accounted for in the analysis. However, this is typically small compared with uncertainty in the genetic associations with the outcome, as variants selected for inclusion in the analysis are typically associated with at least one of the traits at a robust level of statistical significance. The effects of the exposure traits on the outcome are assumed to be linear. This is usually a reasonable assumption, given the small influence of genetic variants on traits, meaning that estimates reflect average causal effects for a small shift in the overall distribution of a trait (Burgess et al., 2014). In our main simulation, all the exposure and outcome traits are continuous. For binary traits, the method can be implemented using genetic association estimates obtained from logistic regression. We have previously shown that multivariable Mendelian randomization methods are still able to make correct inferences in this setting (Burgess & Thompson, 2015; Grant & Burgess, 2021), although the interpretation of estimates is obscured due to non-collapsibility (Burgess & CHD CRP Genetics Collaboration, 2013). The analysis model assumes that variant correlations are the same in all datasets. In the context of the applied example, this may not hold as genetic associations with the exposures were estimated in a Finnish data set, variant associations with the outcome were estimated in European ancestry individuals from the MEGASTROKE consortium, and variant correlations were estimated in European ancestry individuals from UK Biobank. It is most critical that variant correlations are estimated in a similar population group to genetic associations with the outcome, as causal inferences are based on genetic associations with the outcome. We were unable to assess the relevance of the variant correlations estimated in UK Biobank to the MEGASTROKE data set. Finally, estimates are subject to weak instrument bias. Whereas in univariable Mendelian randomization, weak instrument bias in a two-sample setting is towards the null (Pierce & VanderWeele, 2012), in multivariable Mendelian randomization, weak instruments can bias estimates in any direction (Zuber et al., 2020). This is because weak instrument bias is analogous to measurement error, as the genetically-predicted values of the exposures are estimated with uncertainty, which in a multivariable regression model can lead to arbitrary bias (Phillips & Smith, 1991; Thouless, 1939). Hence it is important to balance the inclusion of several genetic variants in the analysis to achieve identification, with the inclusion of only variants strongly associated with the exposures to minimize weak instruments. The optimal balance will depend on the specifics of the analysis (such as the sample size available), but researchers should consider the conditional strength of instruments (via conditional F statistics) as well as the more conventional univariable F statistics (Sanderson et al., 2019). Performance with weak instruments was mixed; mean estimates from the MV-LIML-PCA method were generally less affected than those from the MV-IVW-PCA method, although both methods had slightly elevated Type 1 error rates in one of the scenarios considered. We therefore recommend that both methods are applied when the instruments are weak, and caution is expressed if the methods give divergent results. Overall, we slightly prefer the MV-IVW-PCA method, as the Type 1 error rates from this method were generally slightly lower.

In summary, multivariable cis-Mendelian randomization can be used to disentangle the causal relationships of traits, such as proteins or gene expression measurements, that are influenced by a cluster of correlated genetic variants. The proposed PCA methods provide a compromise between loss of precision resulting from overpruning and numerical instability resulting from underpruning, to allow valid statistical tests that identify the causal traits influencing the outcome.

ACKNOWLEDGMENTS

This study was funded by United Kingdom Research and Innovation Medical Research Council (MC_UU _00002_ 7 and MR/S037675/1) and supported by the National Institute for Health Research Cambridge Biomedical Research Centre (BRC-1215-20014). Fatima Batool and Stephen Burgess are supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (204623 /Z /16/Z). The views expressed are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care. Dr Gill is supported by the British Heart Foundation Centre of Research Excellence (RE/18/4/34215) at Imperial College London and a National Institute for Health Research Clinical Lectureship at St. George's, University of London (CL-2020-16-001).

CONFLICTS OF INTEREST

Dr Gill is employed part-time by Novo Nordisk. The remaining authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The summary statistics for genetic associations with three cytokines that support the findings of this study are available through (Ahola-Olli et al., 2017). The summary statistics for genetic associations with any stroke and cardioembolic stroke are available in MEGASTROKE consortium (Malik et al., 2018) and data were derived from the public domain at [www.megastroke.org]. The variants correlation were accessed in UK Biobank (Sudlow et al., 2015) at [www.ukbiobank.ac.uk]. This study has been conducted using the UK Biobank Resource under Application Number 7439.

ORCID

Fatima Batool http://orcid.org/0000-0002-5375-0628

REFERENCES

Ahola-Olli, A. V., Würtz, P., Havulinna, A. S., Aalto, K., Pitkänen, N., Lehtimäki, T., Kähönen, M., Lyytikäinen, L. P., Raitoharju, E., Seppälä, I., Sarin, A. P., Ripatti, S., Palotie, A., Perola, M., Viikari, J. S., Jalkanen, S., Aksimow, M., Salomaa, V., Salmi, M., Kettunen, J., & Raitakari, O. T. (2017). Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *American Journal of Human Genetics*, *100*(1), 40–50.

Anderson, T., & Rubin, H. (1949). Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics*, *21*(1), 570–582.

Baum, C., Schaffer, M., & Stillman, S. (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal*, *7*(4), 465–506.

Bowden, J., DaveySmith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314.

Bowden, J., Del Greco M. F., Minelli, C., Davey Smith, G., Sheehan, N., & Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, *36*(11), 1783–1802.

Burgess, S., Butterworth, A. S., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*(7), 658–665.

Burgess, S., & CHD CRP Genetics Collaboration. (2013). Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, *32*(27), 4726–4747.

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., Hartwig, F. P., Holmes, M. V., Minelli, C., Relton, C. L., & Theodoratou, E. (2019). Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research*, *4*, 186.

Burgess, S., Davies, N. M., Thompson, S. G., & EPIC-InterAct Consortium (2014). Instrumental variable analysis with a nonlinear exposure-outcome relationship. *Epidemiology*, *25*(6), 877–885.

Burgess, S., Dudbridge, F., & Thompson, S. G. (2015). Re: "Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects." *American Journal of Epidemiology*, *181*(4), 290–291.

Burgess, S., Dudbridge, F., & Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Statistics in Medicine*, *35*(11), 1880–1906.

Burgess, S., Foley, C. N., & Zuber, V. (2018). Inferring causal relationships between risk factors and outcomes from genome-wide association study data. *Annual Review of Genomics and Human Genetics*, *19*, 303–327.

Burgess, S., Thompson, D. J., Rees, J. M., Day, F. R., Perry, J. R., & Ong, K. K. (2017). Dissecting causal pathways using Mendelian randomization with summarized genetic data: Application to age at menarche and risk of breast cancer. *Genetics*, *207*, 481–487.

Burgess, S., & Thompson, S. G. (2015). Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, *181*(4), 251–260.

Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B., & Hopewell, J. C. (2017). Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic Epidemiology*, *41*(8), 714–725.

Carter, A. R., Sanderson, E., Hammerton, G., Richmond, R. C., Davey Smith, G., Heron, J., Taylor, A. E., Davies, N. M., & Howe, L. D. (2021). Mendelian randomisation for mediation analysis: Current methods and challenges for implementation. *European Journal of Epidemiology*, *36*(5), 465–478.

Chao, J. C., & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, *73*(5), 1673–1692.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, *16*(4), 309–330.

Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M., & Brook, R. D. (2015). Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in *NPC1L1*, *HMGCR*, or both: A 2 × 2 factorial Mendelian randomization study. *Journal of the American College of Cardiology*, *65*(15), 1552–1561.

Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S., & Howson, J. M. M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Commununications*, *12*(1), 764.

Georgakis, M. K., De Lemos, J. A., Ayers, C., Wang, B., Björkbacka, H., Pana, T. A.,, Thorand, B., Sun, C., Fani, L., Malik, R., Dupuis, J., Engström, E., Orho-Melander, M., Melander, O., Boekholdt, S. M., Zierer, A., Elhadad, M. A., Koenig, W., Herder, C., … Dichgans, M. (2021). Association of circulating monocyte chemoattractant protein-1 levels with cardiovascular mortality: A meta-analysis of population-based studies. *JAMA Cardiology*, *6*(5), 587–592.

Georgakis, M. K., Gill, D., Rannikmäe, K., Traylor, M., Anderson, C. A., MEGASTROKE consortium of the International Stroke Genetics Consortium (ISGC), Lee, J.-M., Kamatani, Y., Hopewell, J. C., Worrall, B. B., Bernhagen, J., Sudlow, C. L. M., Malik, R., & Dichgans, M. (2019). Genetically determined levels of circulating cytokines and risk of stroke: Role of monocyte chemoattractant protein-1. *Circulation*, *139*(2), 256–268.

Georgakis, M. K., Malik, R., Björkbacka, H., Pana, T. A., Demissie, S., Ayers, C., Elhadad, M. A., Fornage, M., Beiser, A. S., Benjamin, E. J., Boekholdt, S. M., Engström, G., Herder, C., Hoogeveen, R. C., Koenig, W., Melander, O., Orhoelander, M., Schiopu, A., Söderholm, M., … Dichgans, M. (2019). Circulating monocyte chemoattractant protein-1 and risk of stroke: Meta-analysis of population-based studies involving 17 180 individuals. *Circulation Research*, *125*(8), 773–782.

Georgakis, M. K., van der Laan, S. W., Asare, Y., Mekke, J. M., Haitjema, S., Schoneveld, A. H., de Jager, S. C. A., Nurmohamed, N. S., Kroon, J., Stroes, E. S. G., de Kleijn, D. P. V., de Borst, G. J., Maegdefessel, L., Soehnlein, O., Pasterkamp, G., & Dichgans, M. (2021). Monocyte-chemoattractant protein-1 levels in human atherosclerotic lesions associate with plaque vulnerability. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *41*(6), 2038–2048.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5): e1004383.

Gill, D., Georgakis, M. K., Walker, V. M., Schmidt, A. F., Gkatzionis, A., Freitag, D. F., Finan, C., Hingorani, A. D., Howson, J. M. M., Burgess, S., Swerdlow, D. I., Davey Smith, G., Holmes, M. V., Dichgans, M., Scott, R. A., Zheng, J., Psaty, B. M., & Davies, N. M. (2021). Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Research*, *6*, 16.

Grant, A. J., & Burgess, S. (2021). Pleiotropy robust methods for multivariable Mendelian randomization. *Statistics in Medicine*, *40*(26), 5813–5830.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, *50*(4), 1029–1054.

IL6R Genetics Consortium and Emerging Risk Factors Collaboration. (2012). Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *Lancet*, *379*(9822), 1205–1213.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189.

Kalaoja, M., Corbin, L. J., Tan, V. Y., Ahola-Olli, A. V., Havulinna, A. S., Santalahti, K.,, Pitkänen, N., Lehtimäki, T., Lyytikäinen, L.-P., Raitoharju, E., Seppälä, I., Kähönen, M., Ripatti, S., Palotie, A., Perola, M., Viikari, J. S., Viikari, S., Maksimow, M., Salomaa, V., … Timpson, N. J. (2021). The role of inflammatory cytokines as intermediates in the pathway from increased adiposity to disease. *Obesity*, *29*(2), 428–437.

Larsson, S. C., Bäck, M., Rees, J. M., Mason, A. M., & Burgess, S. (2020). Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: A Mendelian randomization study. *European Heart Journal*, *41*(2), 221–226.

Lattka, E., Illig, T., Heinrich, J., & Koletzko, B. (2010). Do FADS genotypes enhance our knowledge about fatty acid related phenotypes? *Clinical Nutrition*, *29*(3), 277–287.

Lawlor, D. A., Harbord, R., Sterne, J., Timpson, N., & DaveySmith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133–1163.

Lawlor, D. A., Tilling, K., & DaveySmith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), 1866.

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A. K., van der Laan, S. W., Gretarsdottir, S., Anderson, C. D., Chong, M., Adams, H. H. H., Ago, T., Almgren, P., Amouyel, P., Ay, H., Bartz, T. M., Benavente, O. R., … Dichgans, M. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics*, *50*(4), 524–537.

Patel, A., Burgess, S., Gill, D., & Newcombe, P. J. (2020). Inference with many correlated weak instruments and summary statistics. *arXiv*, *2005*.01765.

Phillips, A. N., & Smith, G. D. (1991). How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology*, *44*(11), 1223–1231.

Pierce, B., & Burgess, S. (2013). Efficient design for Mendelian randomization studies: Subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology*, *178*(7), 1177–1184.

Pierce, B., & VanderWeele, T. (2012). The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *International Journal of Epidemiology*, *41*(5), 1383–1393.

Porcu, E., Rüeger, S., Lepik, K., Santoni, F. A., Reymond, A., & Kutalik, Z. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, *10*(1), 1–12.

Reijmerink, N. E., Postma, D. S., & Koppelman, G. H. (2010). The candidate gene approach in asthma: What happens with the neighbours? *European Journal of Human Genetics*, *18*(1), 17.

Saleheen, D., Haycock, P. C., Zhao, W., Rasheed, A., Taleb, A., Imran, A., Abbas, S., Majeed, F., Akhtar, S., Qamar, N., Zaman, K. S., Yaqoob, Z., Saghir, T., Rizvi, S. N. H., Memon, A., Mallick, N. H., Ishaq, M., Rasheed, S. Z., Danesh, J., … Memon, F. U. (2017). Apolipoprotein(a) isoform size, lipoprotein(a) concentration, and coronary artery disease: A mendelian randomisation analysis. *Lancet Diabetes & Endocrinology*, *5*(7), 524–533.

Sanderson, E., DaveySmith, G., Windmeijer, F., & Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single sample and two-sample summary data settings. *International Journal of Epidemiology*, *48*(3), 713–727.

Sanderson, E., & Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, *190*(2), 212–221.

Schmidt, A. F., Finan, C., Gordillo-arañón, M., Asselbergs, F. W., Freitag, D. F., Patel, R. S., Tyl, B., Chopade, S., Faraway, R., Zwierzyna, M., & Hingorani, A. D. (2020). Genetic drug target validation using Mendelian randomisation. *Nature Communications*, *11*(1), 3255.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779.

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., … Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*(7708), 73–79.

Thouless, R. H. (1939). The effects of errors of measurement on correlation coefficients. *British Journal of Psychology*, *29*(4), 383–403.

Timms, A. E., Crane, A. M., Sims, A.-M., Cordell, H. J., Bradbury, L. A., Abbott, A., Coyne, M. R. E., Beynon, O., Herzberg, I., Duff, G. W., Calin, A., Cardon, L. R., Wordsworth, B. P., & Brown, M. A. (2004). The interleukin 1 gene cluster contains a major susceptibility locus for ankylosing spondylitis. *The American Journal of Human Genetics*, *75*(4), 587–595.

VanderWeele, T., TchetgenTchetgen, E., Cornelis, M., & Kraft, P. (2014). Methodological challenges in Mendelian randomization. *Epidemiology*, *25*(3), 427–435.

Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H. M., McNeill, M., Smyth, D. J., Niblett, D., Cambien, F., Tiret, L., Todd, J. A., Clayton, D. G., & Blankenberg, S. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics*, *21*(12), 2815–2824.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., & Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*(4), 369-75, S1-3.

Zhu, G., Whyte, M. K., Vestbo, J., Carlsen, K., Carlsen, K. H., Lenney, W., Silverman, M., Helms, P., & Pillai, S. G. (2008). Interleukin 18 receptor 1 gene polymorphisms are associated with asthma. *European Journal of Human Genetics*, *16*(9), 1083–1090.

Zhu, J., Burgess, S., & Grant, A. J. (2022). Bias in multivariable Mendelian randomization studies due to measurement error on exposures. *arXiv*, *2203*.08668.

Zuber, V., Colijn, J. M., Klaver, C., & Burgess, S. (2020). Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nature Communications*, *11*, 29.

Zuber, V., Gill, D., Ala-Korpela, M., Langenberg, C., Butterworth, A., Bottolo, L., & Burgess, S. (2021). High-throughput multivariable Mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease. *International Journal of Epidemiology*, *50*(3), 893–901.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Batool, F., Patel, A., Gill, D., & Burgess, S. (2022). Disentangling the effects of traits with shared clustered genetic predictors using multivariable Mendelian randomization. *Genetic Epidemiology*, 1–15. https://doi.org/10.1002/gepi.22462