Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics

Liza Darrous[1,2,*], Ninon Mounier[1,2,*], Zoltán Kutalik[1,2,3,†]

Abstract

Mendelian Randomisation (MR), an increasingly popular method that estimates the causal effects of risk factors on complex human traits, has seen several extensions that relax its basic assumptions. However, most of these extensions suffer from two major limitations; their under-exploitation of genome-wide markers, and sensitivity to the presence of a heritable confounder of the exposure-outcome relationship. To overcome these limitations, we propose a Latent Heritable Confounder MR (LHC-MR) method applicable to association summary statistics, which estimates bi-directional causal effects, direct heritability, and confounder effects while accounting for sample overlap. We demonstrate that LHC-MR outperforms several existing MR methods in a wide range of simulation settings and apply it to summary statistics of 13 complex traits. Besides several concordant results, LHC-MR unravelled new mechanisms (how being diagnosed for certain diseases might lead to improved lifestyle) and revealed potential false positive findings of standard MR methods (apparent causal effect of body mass index on educational attainment may be driven by a strong ignored confounder). Phenome-wide search to identify LHC-implied heritable confounders showed remarkable agreement between the LHC-estimated causal effects of the latent confounder and those for the potentially identified ones. Finally, LHC-MR naturally decomposes genetic correlation to causal effect-driven and confounder-driven contributions, demonstrating that the genetic correlation between systolic blood pressure and diabetes is predominantly confounder-driven.

[1]University Center for Primary Care and Public Health, University of Lausanne, 1010, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

[3]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, UK

[*]These authors contributed equally to this work.

[†]Correspondence should be addressed to zoltan.kutalik@unil.ch

1 Introduction

Identification of modern day risk factors and quantification of their impact on common diseases is a central question for public health policy makers. Epidemiological studies aim to address this burning issue, but they are most often based on observational data due to its glaring abundance. Despite major methodological advances, a large majority of such studies have inherent limitations and suffer from confounding and reverse causation[1,2]. For these reasons, many of the reported associations found in classical epidemiological studies are mere correlates of disease risk, rather than causal factors directly involved in disease progression. This renders them obsolete when developing public health interventions in a bid to reduce the future burden of a diseases. While well designed and carefully conducted randomised control trials (RCTs) remain the gold standard for causal inference, they are exceedingly expensive, time-consuming, may not be feasible for ethical reasons, and have high failure rates[3,4].

Mendelian randomisation (MR) is an instrumental variable technique used to infer the strength of a causal relationship between a risk factor (X) and an outcome (Y)[5]. To do so, it uses genetic variants (G) as instruments and relies on three major assumptions (see Figure S1): (1) Relevance – G is robustly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome relationship (i.e. the only path between the instrument and the outcome is via the exposure).

The advantage of the MR approach is that for most heritable exposures, dozens (if not hundreds) of genetic instruments are known to date thanks to well-powered genome-wide association studies (GWASs). Each instrument can provide a causal effect estimate, which can be combined with others, by using an inverse variance-weighting (IVW) scheme (e.g. Burgess *et al.*[6]). However, the last assumption is particularly problematic, because genetic variants tend to be pleiotropic, i.e. exert effect on multiple traits independently. Still, it can be shown that if the instrument strength is independent of the direct effect on the outcome (InSIDE assumption) and the direct effects are on average zero, IVW-based methods will still yield consistent estimates. Methods, such as MR Egger[7], produce consistent estimates even if direct effects are allowed to have a non-zero offset. The third assumption can be further reduced to assuming that $> 50\%$ of the instruments (or in terms of their weight) are valid (median-based estimators[8]) or that zero-pleiotropy instruments are the most frequent (mode-based estimators[9]).

The InSIDE assumption (i.e. horizontal pleiotropic effects ($G \to Y$) are independent of the direct effect ($G \to X$)) is reasonable if the pleiotropic path $G \to Y$ does not branch off to X. However, if there is such a branching off, the variable representing the split is a confounder of the $X - Y$ relationship and we fall back on the violation of the second assumption (exchangeability), making it the most problematic. Therefore, in this paper, we extend the standard MR model to incorporate the presence of a latent (unmeasured) heritable confounder (U) and estimate its contribution to traits X and Y, while simultaneously estimating the bi-directional causal effect between the two traits. Standard MR methods are vulnerable to such heritable confounders, as any genetic marker directly associated with the confounder may be selected as an instrument for the exposure. However, such instruments will have a direct effect on the outcome that is correlated to their instrument strength, violating the InSIDE assumption. A very recent approach (CAUSE) proposes a structural equation mixed effect model[10], similar to ours. In the Discussion, we will provide a detailed account of the differences between CAUSE and our method.

The outline of the paper is as follows: first, the extended MR model is introduced and the likelihood function for the observed genome-wide summary statistics (for X and Y) is derived. We then test and compare the method against conventional MR approaches through extensive

simulation settings, including several violations of the model assumptions. Finally, the approach is applied to association summary statistics (based on the UK Biobank and meta-analysis studies) of 13 complex traits to re-assess all pairwise bi-directional causal relationships between them.

2 Methods

2.1 The underlying structural equation model

Let X and Y denote continuous random variables representing two complex traits. Let us assume (for simplicity) that there is one heritable confounder U of these traits. To simplify notation we assume that $E(X) = E(Y) = E(U) = 0$ and $Var(X) = Var(Y) = Var(U) = 1$. The genome-wide sequence data for m markers is denoted by $G = (G_1, G_2, \ldots, G_m)$. The aim of our work is to dissect the effects of the heritable confounding factor U from the bi-directional causal effects of these two traits (X and Y). For this we consider a model (see Figure 1) defined by the following equations.

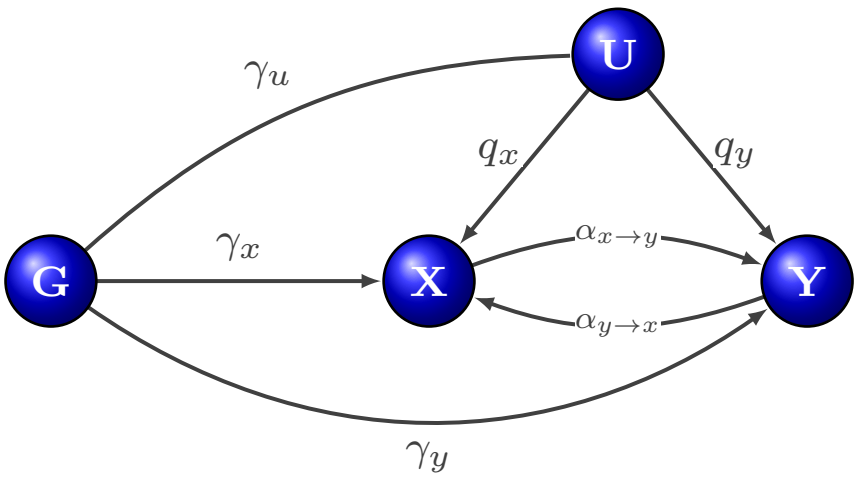

Figure 1: Schematic representation of the extended structural equation model (SEM). X and Y are two complex traits under scrutiny with a latent (heritable) confounder U with causal effects q_x and q_y on them. G represents a genetic instrument, with effects γ_x, γ_y and γ_u, respectively. Traits X and Y have causal effects on each other, which are denoted by $\alpha_{x\to y}$ and $\alpha_{y\to x}$.

$$
\begin{aligned}
X &= q_x \cdot U + \alpha_{y\to x} Y + G \cdot \boldsymbol{\gamma_x} + e_X && \text{with} \quad e_X \sim \mathcal{N}(0, \sigma_x^2) \\
Y &= q_y \cdot U + \alpha_{x\to y} X + G \cdot \boldsymbol{\gamma_y} + e_Y && \text{with} \quad e_Y \sim \mathcal{N}(0, \sigma_y^2) \\
U &= G \cdot \boldsymbol{\gamma_u} + e_U && \text{with} \quad e_U \sim \mathcal{N}(0, \sigma_u^2)
\end{aligned}
$$

where $\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}, \boldsymbol{\gamma_u} \in \mathcal{R}^m$ denote the (true multivariate) direct effect of all m genetic variants on X, Y and U, respectively. Note that the model cannot be represented by classical directed acyclic graphs, as the bidirectional causal effects could form a cycle. It is often proposed[11] to circumvent DAG violation by introducing a temporal aspect (exposure and outcome measured at multiple times), we however, estimate the impact of an infinite loop and model with the equilibrium state (i.e. converged variables).

Note that we do not include in the model reverse causal effects to the confounder ($X \to U$ and $Y \to U$). Let s_x and s_y denote those causal effect of X and Y on U. It is easy to see that by reparameterising the original model to $\alpha'_{x\to y} := \alpha_{x\to y} + s_x \cdot q_y$, $\alpha'_{y\to x} := \alpha_{y\to x} + s_y \cdot q_x$ and $q'_x := q_x/(1-q_x \cdot s_x)$, $q'_y := q_y/(1-q_y \cdot s_y)$, the genetic effects produced by the extended model with

reverse causal effects on U and the simpler model (Figure 1) with the updated parameters are indistinguishable. Thus these extra parameters are not identifiable and the reparameterisation means that $\alpha_{x\to y}$ and $\alpha_{y\to x}$ in our model represent the total causal effects, some of which may be mediated by U.

We model the genetic architecture of these direct effects with a spike-and-slab distribution, assuming that only $0 \leq \pi_x, \pi_y, \pi_u \leq 1$ proportion of the genome have a direct effect on X, Y, U, respectively and these direct effects come from a Gaussian distribution.

$$\begin{aligned}
\boldsymbol{\gamma_x} &\sim \pi_x \cdot \mathcal{N}(0, h_x^2/(m \cdot \pi_x)) + (1-\pi_x) \cdot \mathcal{N}(0,0) \\
\boldsymbol{\gamma_y} &\sim \pi_y \cdot \mathcal{N}(0, h_y^2/(m \cdot \pi_y)) + (1-\pi_y) \cdot \mathcal{N}(0,0) \\
\boldsymbol{\gamma_u} &\sim \pi_u \cdot \mathcal{N}(0, h_u^2/(m \cdot \pi_u)) + (1-\pi_u) \cdot \mathcal{N}(0,0)
\end{aligned}$$

Here, we refer to h_x^2 as the *direct heritability* of X, i.e. independent of the genetic basis of U and Y. Similar notation is adapted for U and Y. Note that when $q_x = 0 and q_y \neq 0$ (or vice versa), this means that there is no confounder U present, but the genetic architecture of Y (or X) can be better described by a three component Gaussian mixture distribution.

We assume that the correlation (across markers) between the direct effects of a genetic variant on X, Y and U is zero, i.e. $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_u}) = cov(\boldsymbol{\gamma_u}, \boldsymbol{\gamma_y}) = 0$. Note that this assumption still allows for a potential correlation between the total effect of G on X ($\gamma_x + \gamma_u \cdot q_x + \gamma_y \cdot \alpha_{y\to x}$) and its horizontal pleiotropic effect on Y ($\gamma_y + \gamma_u \cdot q_y$), but only due to the confounder U and through the reverse causal effect $Y \to X$. As we argued above, this is a reasonable assumption, since the most plausible reason (apart from outcome-dependent sampling) for the violation of the InSIDE assumption may be one or more heritable confounder(s).

For simplicity, we also assume that the set of genetic variants with direct effects on each trait overlap only randomly, i.e. the fraction of the genome directly associated with both X and Y is $\pi_x \cdot \pi_y$, etc. This assumption is in line with recent observation that the bulk of observed pleiotropy can be explained by extreme polygenicity with random overlap between trait loci[12]. Note that uncorrelated effects (e.g. $cov(\boldsymbol{\gamma_x}, \boldsymbol{\gamma_y}) = 0$) do not ensure that the active variant sets overlap randomly, this is a slightly stronger assumption.

2.2 The observed association summary statistics

Let us now assume that we observe univariate association summary statistics for these two traits from two (potentially overlapping) finite samples N_x and N_y of size n_x, n_y, respectively. In the following, we will derive observed summary statistics in sample N_x and then we will repeat the analogous exercise for sample N_y. Let the realisations of X, Y and U be denoted by $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{u} \in \mathcal{R}^{n_x}$. The genome-wide genetic data is represented by $\mathsf{G}_x \in \mathcal{R}^{n_x \times m}$ and the genetic data for a single nucleotide polymorphism (SNP) k tested for association is $\boldsymbol{g_k} \in \mathcal{R}^{n_x}$. Note the distinction between the jth column of G_x, which is the j-th sequence variant, in contrast to g_j, which is the j-th SNP tested for association in the GWAS. We assume that all SNP genotypes have been standardised to have zero mean and unit variance. The association summary statistic

for SNP k of trait X can then be written as

$$\begin{aligned}
\widehat{\beta}_k^x &= \boldsymbol{g}_k' \cdot \boldsymbol{x}/n_x = (q_x \cdot \boldsymbol{g}_k' \cdot \boldsymbol{u} + \alpha_{y\to x} \cdot \boldsymbol{g}_k' \cdot \boldsymbol{y} + \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_x} + \boldsymbol{g}_k' \cdot e_x)/n_x \\
&= q_x \cdot \boldsymbol{g}_k \cdot (\mathsf{G}_x \cdot \boldsymbol{\gamma_u} + e_u)/n_x \\
&+ \alpha_{y\to x} \cdot \boldsymbol{g}_k' \cdot (q_y \cdot (\mathsf{G}_x \cdot \boldsymbol{\gamma_u} + e_u) + \alpha_{x\to y}\boldsymbol{x} + \mathsf{G}_x \cdot \boldsymbol{\gamma_y} + e_y)/n_x \\
&+ (\boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_x} + \boldsymbol{g}_k' \cdot e_x)/n_x \\
&= q_x \cdot \boldsymbol{g}_k \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_u}/n_x + q_x \cdot e_k^u \\
&+ \alpha_{y\to x} q_y \cdot \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_u}/n_x + \alpha_{y\to x} q_y \cdot e_k^u + \alpha_{y\to x}\alpha_{x\to y}\widehat{\beta}_k^x + \alpha_{y\to x} \cdot \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_y}/n_x + \alpha_{y\to x} e_k^y \\
&+ \boldsymbol{g}_k' \cdot \mathsf{G}_x \cdot \boldsymbol{\gamma_x}/n_x + e_k^x
\end{aligned}$$

with $\widehat{\beta}_k^x = \boldsymbol{g}_k' \cdot \boldsymbol{x}/n_x$, $\epsilon_k^u := \boldsymbol{g}_k' \cdot \epsilon_u/n_x \sim \mathcal{N}(0, \sigma_u^2/n_x), \epsilon_k^x := \boldsymbol{g}_k' \cdot \epsilon_x/n_x \sim \mathcal{N}(0, \sigma_x^2/n_x)$ and $\epsilon_k^y := \boldsymbol{g}_k' \cdot \epsilon_y/n_x \sim \mathcal{N}(0, \sigma_y^2/n_x)$. By denoting the linkage disequilibrium (LD) between variant k and all markers in the genome with $\boldsymbol{\rho_k} = \boldsymbol{g}_k' \cdot \mathsf{G}_x/n_x$ we get

$$\begin{aligned}
\widehat{\beta}_k^x &= q_x \cdot \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x} \cdot q_y \cdot \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x} \cdot \alpha_{x\to y} \cdot \widehat{\beta}_k^x + \alpha_{y\to x} \cdot \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_y} + \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_x} \\
&+ (q_x + \alpha_{y\to x} \cdot q_y) \cdot \epsilon_k^u + \epsilon_k^y + \epsilon_k^x \\
&= (\alpha_{y\to x} \cdot q_y + q_x) \cdot \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x} \cdot \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_y} + \boldsymbol{\rho_k'} \cdot \boldsymbol{\gamma_x} + \alpha_{y\to x} \cdot \alpha_{x\to y} \cdot \widehat{\beta}_k^x + \eta_k^x
\end{aligned}$$

with $\eta_k^x := (q_x + \alpha_{y\to x} \cdot q_y) \cdot \epsilon_k^u + \alpha_{y\to x} \cdot \epsilon_k^y + \epsilon_k^x \sim \mathcal{N}(0, i_x/n_x)$, where $i_x := (\alpha_{y\to x} \cdot q_y + q_x)^2\sigma_u^2 + \alpha_{y\to x}\sigma_y^2 + \sigma_x^2$ is equivalent to the LD score regression intercept[13]. Regrouping all $\widehat{\beta}_k^x$ terms on the left side we get

$$\widehat{\beta}_k^x \approx (1 - \alpha_{y\to x} \cdot \alpha_{x\to y}) \cdot \widehat{\beta}_k^x = \boldsymbol{\rho_k'} \cdot ((\alpha_{y\to x} \cdot q_y + q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x} \cdot \boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}) + \eta_k^x$$

We introduced the simplification $(1 - \alpha_{y\to x} \cdot \alpha_{x\to y}) \approx 1$ because in real examples, one of the causal effects is zero, or the product of the bidirectional causal effects is very small.

Similarly, for $\widehat{\beta}_k^y$ estimated in the other sample (N_y) we obtain

$$\widehat{\beta}_k^y \approx \boldsymbol{\rho_k'} \cdot ((\alpha_{x\to y} \cdot q_x + q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y} \cdot \boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}) + \eta_k^y$$

with $\eta_k^y \sim \mathcal{N}(0, i_y/n_y)$, where $i_y := (\alpha_{x\to y} \cdot q_x + q_y)^2\sigma_u^2 + \alpha_{x\to y}\sigma_x^2 + \sigma_y^2$.

We will now partition the genomic variants into eight disjoint groups based on whether they affect (1) only U, (2) only X, (3) only Y, (4) only U and X, (5) only U and Y, (6) only X and Y, (7) U, X and Y, (8) none of them. Due to the independence assumption (i.e. the fraction of the genome associated with both U and X is $\pi_u \cdot \pi_x$) the size of these groups can readily be expressed as a function of the πs, e.g. the proportion of SNPs with effect only on U is $\pi_u \cdot (1 - \pi_x) \cdot (1 - \pi_y)$. If we define $l_k = \boldsymbol{\rho_k'} \cdot \boldsymbol{\rho_k}$ as the LD score[13], the joint distribution of $(\widehat{\beta}_k^x, \widehat{\beta}_k^y)$ can be written as

$$\begin{pmatrix} \widehat{\beta}_k^x \\ \widehat{\beta}_k^y \end{pmatrix} \sim \sum_{i=1}^{8} \omega_i \cdot \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_i\right) \quad \text{where}$$

$$\begin{aligned}
\Omega_1 &= l_k \cdot \Sigma_u + \Sigma_0, & \omega_1 &= \pi_u(1-\pi_x)(1-\pi_y) \\
\Omega_2 &= l_k \cdot \Sigma_x + \Sigma_0, & \omega_2 &= \pi_x(1-\pi_u)(1-\pi_y) \\
\Omega_3 &= l_k \cdot \Sigma_y + \Sigma_0, & \omega_3 &= \pi_y(1-\pi_u)(1-\pi_x) \\
\Omega_4 &= l_k \cdot (\Sigma_u + \Sigma_x) + \Sigma_0, & \omega_4 &= \pi_u\pi_x(1-\pi_y) \\
\Omega_5 &= l_k \cdot (\Sigma_u + \Sigma_y) + \Sigma_0, & \omega_5 &= \pi_u\pi_y(1-\pi_x) \\
\Omega_6 &= l_k \cdot (\Sigma_x + \Sigma_y) + \Sigma_0, & \omega_6 &= \pi_x\pi_y(1-\pi_u) \\
\Omega_7 &= l_k \cdot (\Sigma_u + \Sigma_x + \Sigma_y) + \Sigma_0, & \omega_7 &= \pi_u\pi_x\pi_y \\
\Omega_8 &= \Sigma_0, & \omega_8 &= (1-\pi_u)(1-\pi_x)(1-\pi_y)
\end{aligned}$$

with

$$
\begin{aligned}
\Sigma_u &= \frac{h_u^2}{m \cdot \pi_u} \begin{bmatrix} (\alpha_{y\to x} \cdot q_y + q_x)^2 & (\alpha_{y\to x} \cdot q_y + q_x) \cdot (\alpha_{x\to y} \cdot q_x + q_y) \\ (\alpha_{y\to x} \cdot q_y + q_x) \cdot (\alpha_{x\to y} \cdot q_x + q_y) & (\alpha_{x\to y} \cdot q_x + q_y)^2 \end{bmatrix} \\
\Sigma_x &= \frac{h_x^2}{m \cdot \pi_x} \begin{bmatrix} 1 & \alpha_{x\to y} \\ \alpha_{x\to y} & \alpha_{x\to y}^2 \end{bmatrix} \\
\Sigma_y &= \frac{h_y^2}{m \cdot \pi_y} \begin{bmatrix} \alpha_{y\to x}^2 & \alpha_{y\to x} \\ \alpha_{y\to x} & 1 \end{bmatrix} \\
\Sigma_0 &= \begin{bmatrix} i_x/n_x & \rho_{x\cap y} \\ \rho_{x\cap y} & i_y/n_y \end{bmatrix}
\end{aligned}
$$

where h_x^2, h_y^2, h_u^2 are the direct heritabilities of X, Y and U, respectively. Parameter $\rho_{x\cap y} := \frac{\rho_{x,y} \cdot n_{x\cap y}}{n_x \cdot n_y}$ represents the cross-trait LD-score intercept[14], where $\rho_{x,y} = corr(X, Y)$ and $n_{x\cap y}$ is the size of the sample overlap between the two studies based on which $\widehat{\beta}_k^x$ and $\widehat{\beta}_k^y$ were calculated. Note that we can define $t_x := q_x \cdot h_u$ and $t_y := q_y \cdot h_u$ to reduce the number of parameters since the formula includes only the product of q_x or q_y and h_u, hence individually they are not identifiable. Parameters $\{n_x, n_y, m, \boldsymbol{l}\}$ are known and the other 12 $\boldsymbol{\theta} = \{i_x, i_y, \pi_u, \pi_x, \pi_y, h_x^2, h_y^2, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, \rho_{x\cap y}\}$ are to be estimated from the observed association summary statistics. It is key to note that our approach does not aim to estimate individual (direct or indirect) SNP effects, as these are handled as random effects. By replacing U with $-U$ we swap the signs of both t_x and t_y, therefore these parameters are unique only if the sign of one of them is fixed. Therefore, we will have the following restrictions on the parameter ranges: π_u, π_x, π_y are in $[0, 1]$, h_x^2, h_y^2, t_x are in $[0, 1]$, $t_y, \alpha_{x\to y}, \alpha_{y\to x}$ are in $[-1, 1]$ and $\rho_{x\cap y} \in [-1/\max(n_x, n_y), 1/\max(n_x, n_y)]$.

The likelihood function is therefore defined as a mixture of eight Gaussian probability density functions (ϕ)

$$
\mathcal{L}\left(\boldsymbol{\theta} \middle| \begin{pmatrix} \widehat{\beta}_k^x \\ \widehat{\beta}_k^y \end{pmatrix}\right) \propto \sum_{i=1}^{8} \omega_i \cdot \phi\left(\begin{pmatrix} \widehat{\beta}_k^x \\ \widehat{\beta}_k^y \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_i\right)
$$

Since different SNPs are correlated we have to estimate the over-counting of each SNP. We choose the same strategy as LD score regression[13] and weigh each SNP by the inverse of its restricted LD score, i.e. as $w_k = 1/\sum_{j=1}^{K} r_{jk}^2$, where $r_{jk} = (\boldsymbol{\rho_k})_j$ is the correlation between SNPs k and j. Namely, the likelihood function of the full set of association summary statistics for K SNPs is defined as

$$
\mathcal{L}\left(\boldsymbol{\theta} \middle| \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{\boldsymbol{x}} \\ \widehat{\boldsymbol{\beta}}^{\boldsymbol{y}} \end{pmatrix}\right) \propto \prod_{k=1}^{K} \left(\sum_{i=1}^{8} \omega_i \cdot \phi\left(\begin{pmatrix} \widehat{\beta}_k^x \\ \widehat{\beta}_k^y \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_i\right)\right)^{w_k} \tag{1}
$$

Our method, termed *Latent Heritable Confounder Mendelian Randomisation (LHC-MR)*, maximises this likelihood function to obtain the maximum likelihood estimate (MLE) and uses the likelihood-ratio test (LRT) to establish the P-value corresponding to each estimated parameter. Assuming that the estimated parameter comes from a Gaussian distribution, the standard error (SE) of the estimator can be computed based on the P-value and the parameter estimate ($SE(\delta) = -|\widehat{\delta}|/\Phi^{-1}(P/2)$). Due to the complexity of the likelihood surface, we initialise the maximisation using 100 different random starting points from a uniform distribution with the parameter-specific ranges mentioned above and chose parameters corresponding to the largest likelihood of the 100 runs. Running time depends on the number of iterations during the maximisation procedure, and is linear with respect to the number of SNPs. It takes ~ 0.35 CPU-minute to fit the complete model to 50,000 SNPs with a single starting point.

Given the particular nature of the underlying directed graph, in certain situations we observed two global optima. The reason for this is the difficulty in distinguishing the ratio of the confounder effects on Y and X from the causal effect of X on Y, as illustrated in Figure S2 by the slopes belonging to different SNP-clusters. The dark blue cluster of SNPs are not in violation of any of the MR assumptions, hence their slope represents the true causal effect of X on Y, while the red cluster of SNPs are those associated to the confounder and their slope represents the ratio of the confounder effects on the two traits. In such situations we keep track of both parameter estimates maximising the likelihood function.

2.3 Decomposition of genetic correlation

Given the starting equations for X and Y we can calculate their genetic correlation. Denoting the total (multivariate) genetic effect for X and Y as δ_x and δ_Y, we can express them as follows

$$\begin{aligned}\delta_x &= q_x \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\delta_y + \boldsymbol{\gamma_x} \\ \delta_y &= q_y \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\delta_x + \boldsymbol{\gamma_y}\end{aligned}$$

Substituting the second equation to the first yields

$$\begin{aligned}\delta_x &= q_x \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}(q_y \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\delta_x + \boldsymbol{\gamma_y}) + \boldsymbol{\gamma_x} \\ &= (q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + (\alpha_{y\to x}\alpha_{x\to y})\delta_x + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x} \\ &= \left((q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})\end{aligned}$$

Similarly,

$$\delta_y = \left((q_y + \alpha_{x\to y}q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})$$

Thus the genetic covariance is

$$\begin{aligned}E[\delta_x \cdot \delta_y] &= \left((q_x + \alpha_{y\to x}q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x}\boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}\right)\left((q_y + \alpha_{x\to y}q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y}\boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ &= \left((q_x + \alpha_{y\to x}q_y)(q_y + \alpha_{x\to y}q_x)h_u^2 + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ &= \left((t_x + \alpha_{y\to x}t_y)(t_y + \alpha_{x\to y}t_x) + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})^2\end{aligned}$$

and the heritabilities are

$$\begin{aligned}E[\delta_x^2] &= \left((t_x + \alpha_{y\to x}t_y)^2 + \alpha_{y\to x}^2 h_y^2 + h_x^2\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})^2 \\ E[\delta_y^2] &= \left((t_y + \alpha_{x\to y}t_x)^2 + \alpha_{x\to y}^2 h_x^2 + h_y^2\right)/(1 - \alpha_{y\to x}\alpha_{x\to y})^2\end{aligned}$$

Therefore the genetic correlation takes the form

$$corr(\delta_x, \delta_y) = \frac{(t_x + \alpha_{y\to x}t_y)(t_y + \alpha_{x\to y}t_x) + \alpha_{y\to x}h_y^2 + \alpha_{x\to y}h_x^2}{\sqrt{\left((t_x + \alpha_{y\to x}t_y)^2 + \alpha_{y\to x}^2 h_y^2 + h_x^2\right)\left((t_y + \alpha_{x\to y}t_x)^2 + \alpha_{x\to y}^2 h_x^2 + h_y^2\right)}} \quad (2)$$

These values can be compared to those obtained by LD score regression (see Results).

2.4 Simulation settings

First, we tested LHC-MR using realistic parameter settings with mild violation of the classical MR assumptions. These standard parameter settings consisted of simulating $m = 50{,}000$ SNPs for two non-overlapping cohorts of equal size (for simplicity) of $n_x = n_y = 500{,}000$ for each trait. X, Y and U were simulated with moderate polygenicity ($\pi_x = 0.1, \pi_y = 0.15, \pi_u = 0.05$), and considerable direct heritability ($h_x^2 = 0.25, h_y^2 = 0.2, h_u^2 = 0.3$). U had a confounding effect

on the two traits as such, $q_x = 0.3, q_y = 0.2$ (resulting in $t_X = 0.16, t_Y = 0.11$), and X had a direct causal effect on Y ($\alpha_{x\to y} = 0.3$), while a reverse causal effect from Y to X was set to null. It is important to note that for the standard parameter setting and each different parameter setting that we tested, we replicated the estimation on 100 different data generations, where each generation underwent a likelihood maximisation of Eq. 1 and produced estimated parameters corresponding to the highest likelihood (schema in Figure S3).

In following simulations, we changed various parameters of these standard settings to test the robustness of the method. We explored how reduced sample sizes ($n_x = n_y$ ranging between 10,000 and 500,000) or differences in sample sizes ($(n_x, n_y) =$ (50,000, 500,000) and $(n_x, n_y) =$ (500,000, 50,000)) influence causal effect estimates of LHC-MR and other MR methods. Next, to ensure that our estimates are not biased if only a fraction of the summary statistics are observed, we randomly masked 90% of the available summary statistics (45,000 out of 50,000). We also changed the proportion of SNPs with non-zero effect on traits X, Y and U (π_x, π_y and π_u) to $35\%, 60\%, 90\%$ and 100%. In addition, we tried to create extremely unfavourable conditions for all MR analyses by increasing the confounding effects. This can be done in two ways: (i) increasing q_x and/or q_y ($q_x = 0.75, q_y = 0.2$), or (ii) increasing the polygenicity of U ($\pi_u = 0.35$).

Finally, we explored various violations of the assumptions of our model (see Section 2). First, we breached the assumption that the non-zero effects come from a Gaussian distribution. By design, the first three moments of the direct effects are fixed: they have zero mean, their variance is defined by the direct heritabilities and they must have zero skewness because the effect size distribution has to be symmetrical. Therefore, to violate the normality assumption, we varied the kurtosis (2, 3, 5, 10 and 15) of the distribution drawn from the Pearson's distribution family. Next, we also simulated data in the presence of two heritable confounders (while still fitting the model with only one U) once with concordant ($t_x^{(1)} = 0.16, t_y^{(1)} = 0.11, t_x^{(2)} = 0.22, t_y^{(2)} = 0.16$) and once with discordant ($t_x^{(1)} = 0.16, t_y^{(1)} = 0.11, t_x^{(2)} = 0.22, t_y^{(2)} = -0.16$) causal effects on X and Y.

2.5 Application to real summary statistics

Once we demonstrated favourable performance of our method on simulated data, we went on to apply LHC-MR to summary statics obtained from the UK Biobank and other meta-analytic studies (Table S2) in order to estimate pairwise bi-directional (univariate) causal effect between 13 complex traits. The traits varied between conventional risk factors such as low education, high body mass index (BMI), dislipidemia and outcomes including diabetes and myocardial infarction among others. For each summary association statistic, SNPs with imputation quality greater than 0.99, and minor allele frequency (MAF) greater than 0.005 were selected. Moreover, SNPs found within the human leukocyte antigen (HLA) region on chromosome 6 were removed due to the abundance of SNPs associated with autoimmune and infectious diseases as well as the complicated LD structure present in that region.

In order to perform LHC-MR between trait pairs, a set of overlapping SNPs was used as input for each pair. The effects of these overlapping SNPs were then aligned to the same effect allele in both traits. In order to decrease computation time (while only minimally reducing power), we selected every 10th QC filtered SNP as input for the analysis. The LHC-MR procedure was run for each pair of traits 100 times, each using a different set of randomly generated starting points within the ranges of their respective 12 parameters. For the optimisation of the likelihood function (Eq. 1), we used the R function 'optim' from the 'stats' R package[15].

Once we fitted this *complete* model estimating 12 parameters $\{i_x, i_y, \pi_u, \pi_x, \pi_y, h_x^2, h_y^2, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, \rho_{x\cap y}\}$, we then fitted nested models obtained by removing each of the following parameters $\{\alpha_{x\to y}, \alpha_{y\to x}, t_X, t_Y, \{t_X, t_Y, \pi_u\}\}$ at a time. To further speed up computation, for

nested sub-models we used the parent model parameters as starting points. The maximum likelihood value of these five nested models was then compared against that of the complete/parent model using the LRT, and the corresponding P-value of each of the above-mentioned parameters was obtained. We first selected the sub-model with the highest ML value. If the LRT comparing the complete and best nested model showed a non-significant difference between them (at 5% family-wise error rate (FWER)), we chose the best nested model as basis for further investigations. We then iteratively repeated the above procedure (comparing the focal model with all its nested sub-models) until the best nested model is identified. This procedure was repeated until a final model was found (i) which was not significantly worse than any of its parent models (i.e. LRT $P \geq 0.05/(2*\text{number of comparison pairs})$) and significantly better (LRT $P < 0.05/(2*\text{number of comparison pairs})$) than all of its nested sub-models.

We first took 4,773,627 SNPs with info (imputation certainty measure) ≥ 0.99 present in the association summary files from the second round of GWAS by the Neale lab[16]. This set was restricted to 4,650,107 common, high-quality SNPs, defined as being present in both UK10K and UK Biobank, having MAF$>$ 1% in both data sets, non-significant ($P_{diff} > 0.05$) allele frequency difference between UK Biobank and UK10K and residing outside the HLA region (chr6:28.5-33.5Mb). For these SNPs, LD scores and regression weights were computed based on 3781 individuals from the UK10K study[17]. For the LD score calculation we computed squared correlations between the target SNP and all sequence variants within its 2 Mb neighbourhood with MAF$\geq$ 0.5% (in the UK10K). For the regression weight calculation, we restricted the neighbouring variants to those part of the 4,650,107 well-imputed SNPs. The procedure may be sub-optimal for summary statistics not coming from the UK Biobank, but we have shown[18] that estimating LD in a ten-times larger data set (UK10K) outweighs the benefit of using smaller, but possibly better-matched European panel (1000 Genomes[19]).

2.6 Comparison against conventional MR and other methods

We compare the causal parameter estimate of the LHC-MR method to those of five conventional MR approaches (MR Egger, weighted median, IVW, mode MR, and weighted mode MR) using a Z-test[20]. The 'TwoSampleMR' R package[21] was used to get the causal estimates for all the pairwise traits as well as their standard errors from the above-mentioned MR methods. The same set of genome-wide SNPs that were used by LHC-MR, were inputted as a starting point for the package. This comprised the selection of genome-wide significant SNPs associated with the exposure (absolute P-value $< 5 \times 10^{-8}$), and removing SNPs that were significantly ($P < 10^{-3}$) more strongly associated with the exposure than the outcome. The default package settings for the clumping of SNPs ($r^2 = 0.001$) were used and the analysis was run with no further changes. We used the Cochran's Q-test to measure the agreement between conventional MR approaches, and comparisons were performed only when the five different MR methods yielded homogeneous causal effect estimations. Note that since the estimates are somewhat correlated, the Cochran test is permissive and eliminates trait pairs with very strong discordance between standard MR methods.

For the pairs that had a statistically significant (at 5% FWER) difference between LHC-MR and at least two standard MR methods (i.e. Z-test $P < 0.05/\text{number of comparison pairs}$), we used a restricted likelihood for LHC-MR allowing neither confounder nor reverse causal effect in order to mimic standard MR assumptions. The LHC causal estimates from this no-confounder-no-reverse-causality model where then again compared against the estimates of all five standard MR methods. We further tested the agreement between the significance of our estimates and that of standard MR methods, with the focus being on finding differences in statistical conclusions regarding causal effect sizes.

3 Results

3.1 Overview of the method

We fitted a 12-parameter structural equation model (SEM) (Figure 1) to genome-wide summary statistics for two studied complex traits in order to estimate bi-directional causal effects between them (for details see Methods). Additional model parameters represent direct and indirect heritabilities for X and Y, cross-trait and individual trait LD score intercepts and the polygenicity for X, Y and U. All SNPs associated with the heritable confounder (U) are indirectly associated to X and Y with effects that are proportional (ratio $t_x : t_y$). SNPs that are directly associated with X (and not with U) are also associated with Y with proportional effects (ratio $1 : \alpha_{x\to y}$). Finally, SNPs that are directly Y-associated are also X-associated with a proportionality ratio of $1 : \alpha_{y\to x}$. These three groups of SNPs are illustrated on the β_x-vs-β_y scatter plot (Figure S2). In simple terms, the aim of our method is to identify the different clusters, estimate the slopes and distinguish which corresponds to the causal- and confounder effects. In this paper, we focus on the properties of the maximum likelihood estimates (and their variances) for the bi-directional causal effects arising from our SEM.

3.2 Simulation results

We started off with a realistic simulation setting of 50,000 SNPs and 500,000 samples for both traits. Traits X, Y and confounder U had average polygenicity ($\pi_x = 0.1, \pi_y = 0.15, \pi_u = 0.05$), with substantial direct heritability for X and Y ($h_x^2 = 0.25, h_y^2 = 0.2$), mild confounding ($t_X = 0.16, t_Y = 0.11$) and causal effect from X to Y ($\alpha_{x\to y} = 0.3, \alpha_{y\to x} = 0$). Note that with these settings, SNPs associated with U would violate the InSIDE assumption but might still be used by conventional MR methods. Under these settings, all methods worked reasonably well, causal effect estimates $\widehat{\alpha}_{x\to y}$ were minimally biased by most conventional MR methods, only MR Egger being noticeably downward biased (bias = -0.05). However, not only was our method the least biased, but it also had four times lower root-mean-square error (RMSE) and ten times lower variance (Figure 2 panel a) thanks to the fact that it efficiently exploits all SNPs genome-wide.

When we lowered the sample sizes ($n_x = n_y$) from 500,000 down to 10,000, as expected, most MR methods became heavily downward biased, because the exposure effect estimates suffer from winner's curse and weak instrument bias. Our method, on the other hand remained the least biased (Figure 3) at the cost of (almost ten-fold) higher variance and 50% increase in RMSE (Table S1) compared to IVW. Furthermore, unequal sample sizes for the two traits showed an underestimation of the causal effects for all MR methods, while LHC remained the most accurate in the case where n_x (50,000) was smaller than n_y (500,000). However, in the inverse case where n_x was larger in size, all MR methods showed correct estimation with high variability, similar to LHC-MR which had a few outliers (Figure S4). We also observed that estimation bias and variance did not change noticeably when 90% of the summary statistics were masked (bias of 2×10^{-5} for non-masked vs. 2×10^{-4} for masked SNPs) (Figure S5).

Reassuringly, our LHC-MR method is not influenced by the extent of polygenicity of X, Y or U either (Figure 2 panel c, Figure S6) and it failed to produce meaningful estimates only in the case of the unlikely scenario where all SNPs genome-wide influence all three traits (see Figure S7). Increasing the proportion of confounder-associated SNPs to 35% only seems to (slightly) bias the weighted median and mode estimators, while other methods (including ours) are left unaffected (Figure S8). In this case our method has the lowest bias and variance (bias $= 5.4 \times 10^{-4}$, variance $= 1.3 \times 10^{-5}$). Further increasing the proportion of confounder associated SNPs to $60\%, 90\%$ and even 100% does not seem to noticeably change the above results (see Figure S6 panel b).

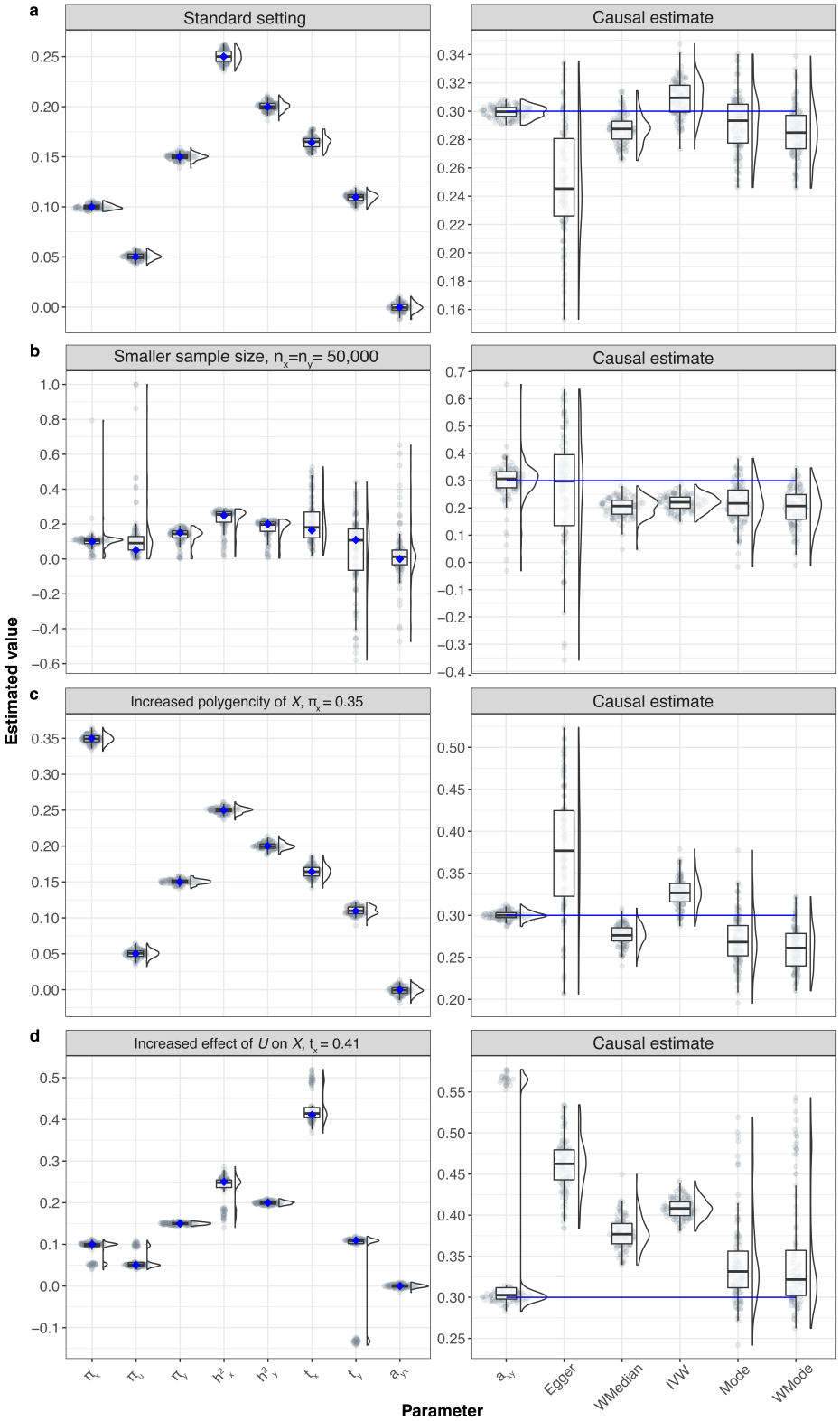

Figure 2: Raincloud boxplots[22] representing the distribution of parameter estimates from 100 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. Panel a represents the estimation under standard settings ($\pi_x = 0.1, \pi_y = 0.15, \pi_u = 0.05, h^2_x = 0.25, h^2_y = 0.2, h^2_u = 0.3, t_X = 0.16, t_Y = 0.11$). Panel b explores smaller sample sizes with $n_x = n_y = 50{,}000$. Panel c introduces an increased polygenicity for X in the form of higher $\pi_x = 0.35$. Panel d shows the increased effect of U on X through $t_X = 0.41$ instead of the basic setting $t_X = 0.16$, and the resulting bimodality.

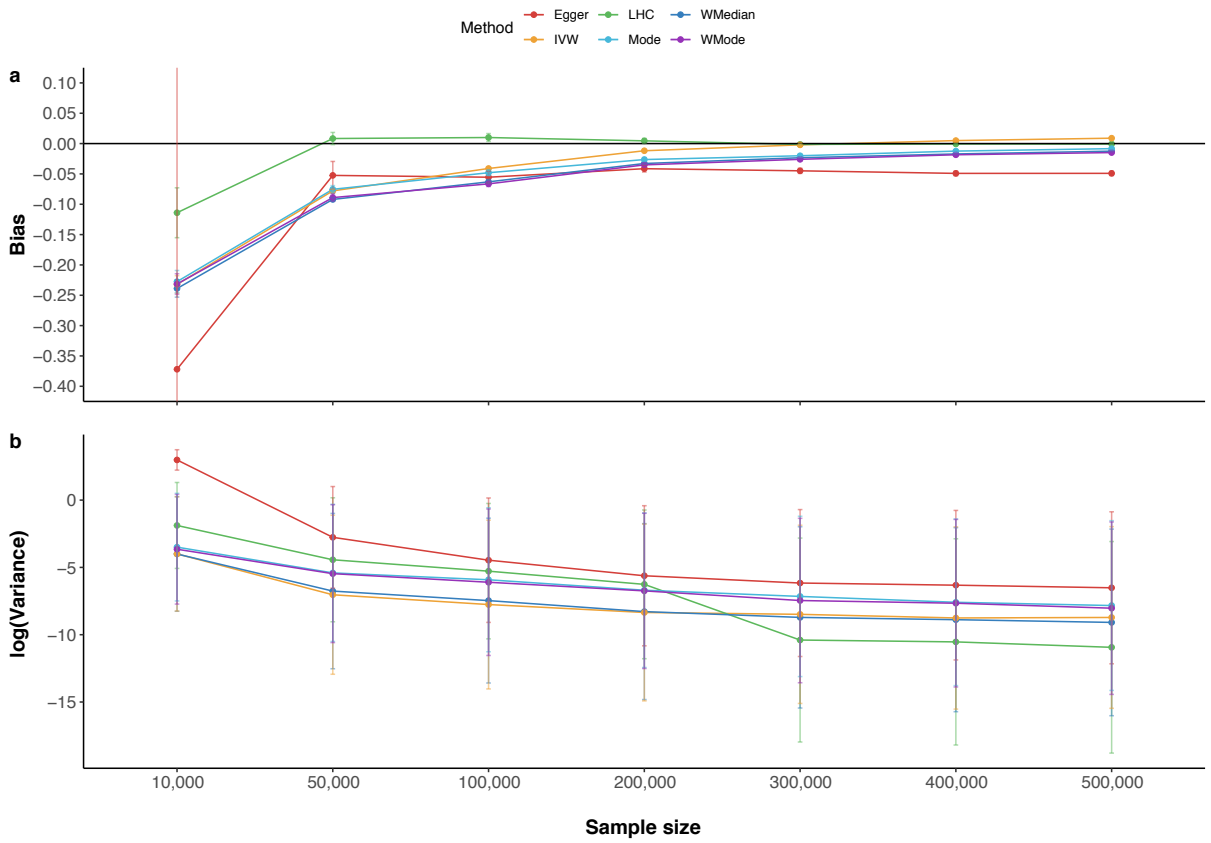

Figure 3: Line plots representing the bias and log variance of the estimated causal parameter, $\alpha_{x \to y}$, using the various MR methods against LHC-MR for different sample size ranging from 10,000 to 500,000 with the standard error reported.

Drastically increasing the indirect genetic effects, by intensifying the contribution of the confounder to X and Y ($t_x = 0.41, t_y = 0.11$), led to a bimodal likelihood surface with two equally favoured local (and global) maxima (Figure 2 panel d). In such situations, the LHC-MR method proposes two equally likely underlying models with different causal effect estimates: $\widehat{\alpha}_{x\to y} = 0.57$ ($\text{Var}(\widehat{\alpha}_{x\to y}) = 3.73\times10^{-5}$) and $\widehat{\alpha}_{x\to y} = 0.30$ ($\text{Var}(\widehat{\alpha}_{x\to y}) = 3.27\times10^{-5}$)). The reason of which is the inability to distinguish between the true causal effect and the ratio of the confounder effects on the exposure and outcome, as seen in Figure S2. Conventional MR methods in this case yield a causal effect estimate in-between the two optima of LHC-MR, with MR Egger having the most extreme relative bias of $\approx 50\%$.

Finally, we explored how sensitive our method is to different violations of our modelling assumptions. First, we simulated summary statistics when the underlying non-zero effects come from a non-Gaussian distribution. We observed that although underlying distributions with excess kurtosis result in downward bias, its extent is minimal. On the other hand, leptokurtic effect size distributions yielded slightly more pronounced upward bias, while still staying below 3% relative bias (Figure 4). Interestingly, other MR methods also exhibit kurtosis-dependent bias with a much more pronounced opposite trend.

Admittedly, the most fundamental assumption of our model is the presence of a single confounder. For this reason, we simulated summary statistics, where the $X - Y$ relationship has two confounders, U_1 and U_2. When the causal effects of these confounders (q_x, q_y) agreed in sign between the confounders, the corresponding causal effects were extremely well estimated in comparison to conventional MR methods (bias $= -1.3 \times 10^{-4}$). However, when the signs were opposite ($q_x^{(1)} = 0.3, q_y^{(1)} = 0.2$ for U_1 and $q_x^{(2)} = 0.4, q_y^{(2)} = -0.3$ for U_2), the causal effect exhibited a bimodal distribution in its estimation, with one peak around the true value, and another one underestimating the causal effect (Figure S9). Conventional MR methods in this case almost all underestimated the causal effect with the exception of MR Egger that had the lowest bias (bias $= 8.4 \times 10^{-3}$).

3.3 Application to association summary statistics of complex traits

We applied our LHC and other MR methods to estimate all pairwise causal effects between 13 complex traits. These results are presented as a heatmap in Figure 5 and as a table in Table XX (LHC-MR_13PairResults.xlsx). Our comparison of the results obtained by LHC-MR and standard MR methods is detailed below and represented in Figure S10. In summary, for 148 out of 156 trait pairs, causal estimates were not significantly discordant between the five different standard MR methods. Thirty-two out of 148 pairs were not ideal for LHC-MR either due to low estimated direct heritability (< 0.01) (i.e. not enough instrument strength) or an estimated bimodal causal effect (i.e. no unique estimate). For 113 out of those 116 comparable trait pairs, our LHC-MR causal effect estimates were concordant (not significantly different) with at least 2 out of 5 standard MR methods' estimates. Only three pairs disagreed with 4 MR methods, 2 other pairs were discordant with 3 MR methods, and 3 pairs had conflicting estimates between LHC and 2 MR methods (discussed later on).

LHC-MR confirmed many previous findings, such as increased BMI leading to elevated blood pressure, diabetes mellitus (DM), myocardial infarction (MI). Furthermore, we confirmed previous results[23] that diabetes increases SBP ($\widehat{\alpha_{x\to y}} = 0.17(P = 2.89 \times 10^{-5})$).

Interestingly, it revealed that higher BMI increases smoking intensity, concordant with other studies[24,25]. It has also shown the protective effect of education against a wide range of diseases (coronary artery disease (CAD), MI, obesity). Our results reveal that many diseases lead to

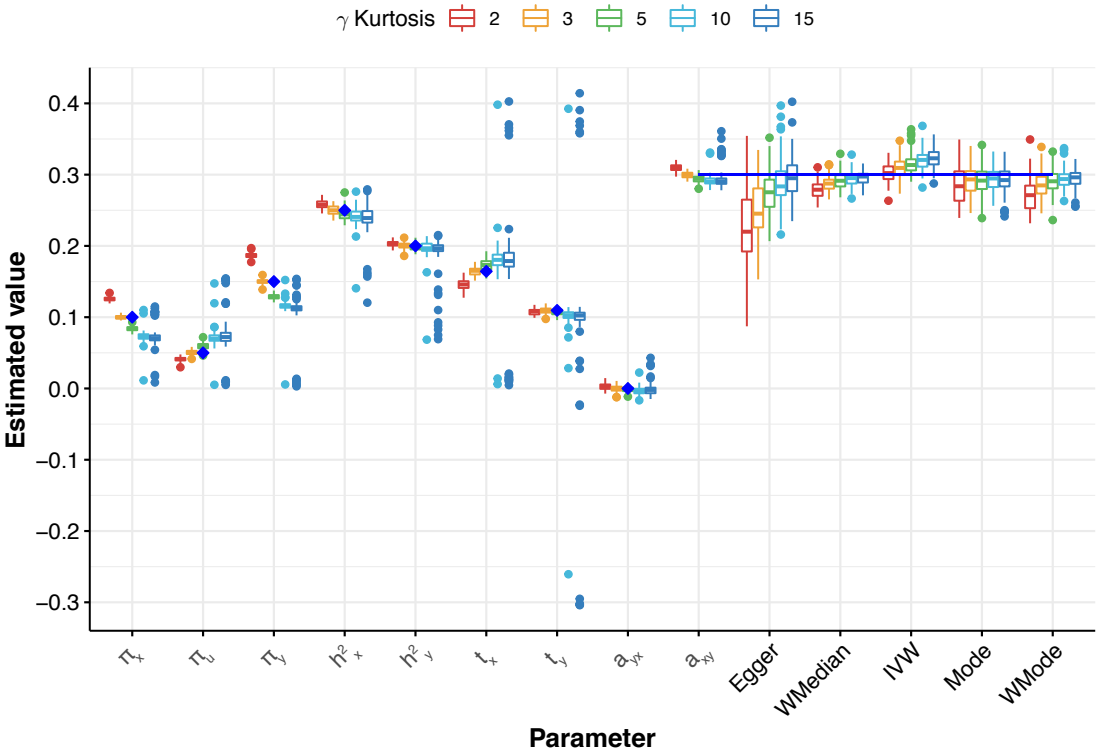

Figure 4: Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. The different coloured boxplots represent the underlying non-normal distribution used in the simulation of the three $\gamma_x, \gamma_x, \gamma_u$ vectors associated to their respective traits. The Pearson distributions had the same 0 mean and skewness, however their kurtosis ranged between 2 and 15, including the kurtosis of 3, which corresponds to a normal distribution assumed by our model.

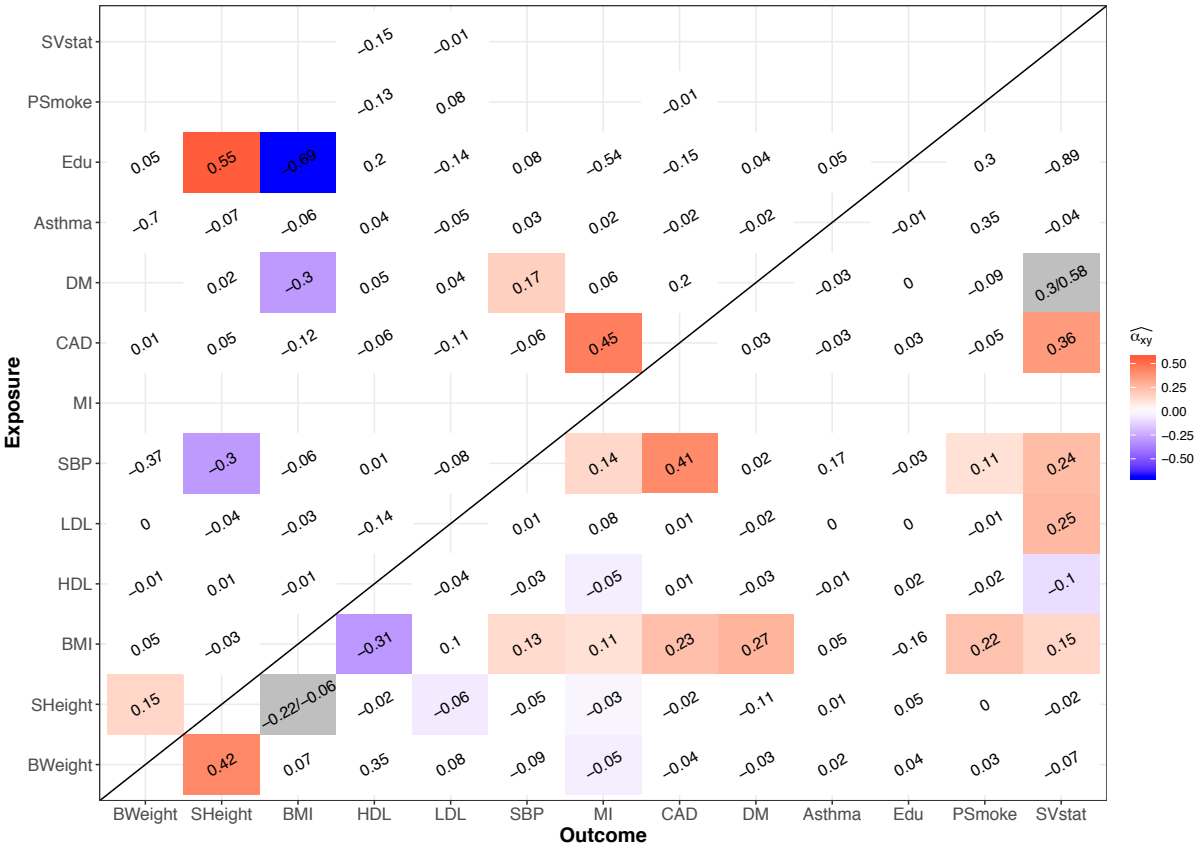

Figure 5: Heatmap representing the bidirectional causal relationship between the 13 UK Biobank traits. The causal effect estimates in coloured tiles all have a significant p-values surviving Bonferroni multiple testing correction with a threshold of 2.4×10^{-4}. When two distinct (global) maxima exist we report the means of each peak in that distribution in the grey boxes (detailed in Figures S11 and S12). Furthermore, we did not report an estimated causal effects for exposures with estimated direct heritability less than 1%. White tiles show an absence of a significant casual effect estimate.
Abbreviations: BMI: Body Mass Index, BWeight: Birth Weight, DM: Diabetes Mellitus, Edu: Years of Education, LDL: Low Density Lipoprotein, MI: Myocardial Infarction, PSmoke: Cigarettes Previously Smoked, SBP: Systolic Blood Pressure, SHeight: Standing Height, SVstat: Medication Simvastatin

lower BMI, probably reflecting lifestyle change recommendations by medical doctors upon the disease diagnosis. In the same vein, statin use is greatly increased when being predisposed to / diagnosed with CAD, (systolic) hypertension, dislipidemia, and diabetes.

Note-worthy, effects of height on SBP and DM has been shown in large MR studies[26]. LHC-MR, while agreeing with these claims in trend, did not find significant evidence to support them and indicates the existence of a confounder with causal effects $-0.51(P = 1.33 \times 10^{-234})$ and $0.02(P = 6.84 \times 10^{-72})$ on height and SBP. A similar confounder was detected for the height-DM relationship with causal effects $-0.05(P = 4.41 \times 10^{-53})$ and $0.52(P = 2.38 \times 10^{-289})$ on height and DM. These relationships require further scrutiny as it has been proposed that these causal effects may be mediated through other traits (e.g. lung function and BMI).

Due to the shared genetics between parents and offspring, some causal effects suggested by MR may represent the effect of parental exposure to offspring outcome[27]. Such examples include the impact of height on birth weight or the height-increasing effects of higher education or low SBP. Reassuringly, no adult trait seems to causally impact the years spent in education. This is not the case for IVW, which (erroneously) indicates that higher BMI decreases education level by 17% ($P = 1.83 \times 10^{-25}$).

A systematic comparison between IVW and LHC-MR has shown generally good agreement between the two methods, which is illustrated in Figure 6. To identify discrepancies between our causal estimates and those of the standard MR results, we grouped the estimates into three categories, either non-significant p-value, significant with an agreeing sign for the causal estimate, or significant with a disagreeing sign. The diagonal (seen in Figure 6) representing the agreement in significance status and sign between the two methods, is heavily populated. On the other hand, six pairs have causal links that are significantly non-zero according to LHC-MR, but are non-significant by IVW, while the opposite is true for 18 pairs. We believe that many of these 18 pairs may be false positives, since only 7 have been confirmed by more robust MR methods, such as weighted mode. Further comparisons of significance between LHC-MR estimates and the remaining standard MR methods can be found in Figure S14. Only one pair (LDL's effect on HDL) differs in sign while being significant in both IVW and LHC-MR, which is probably due to the fact that IVW ignored the presence of a strong confounder(s), which we will further evidence below.

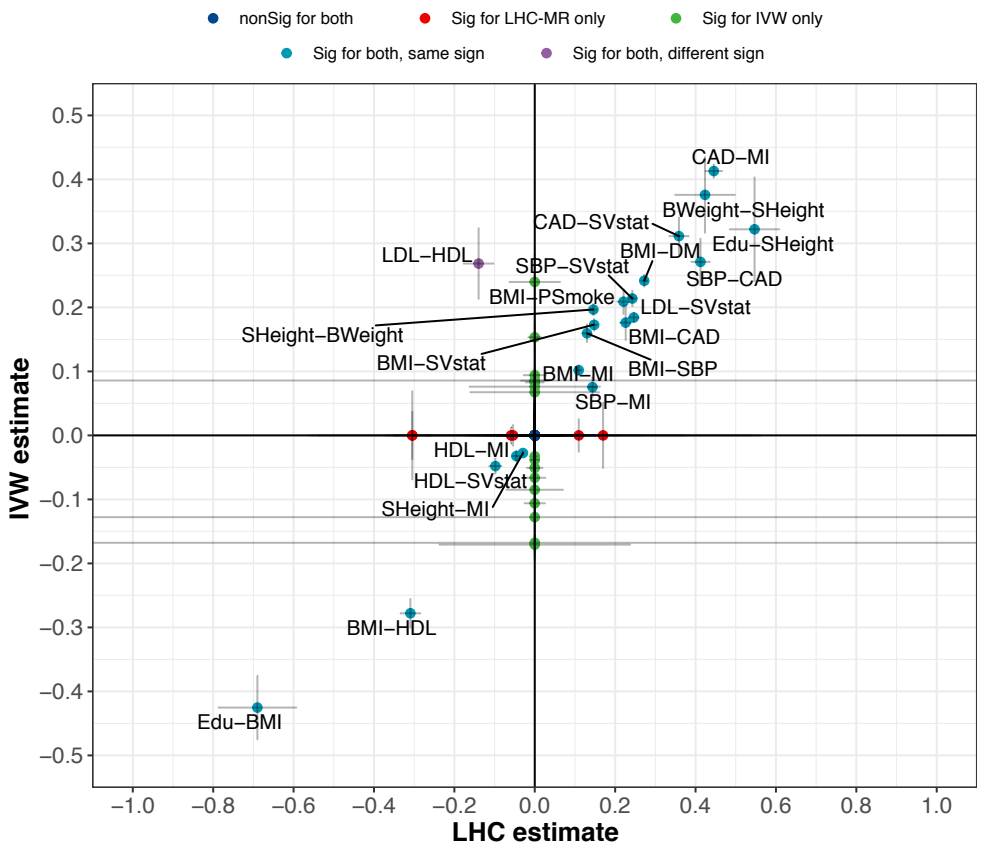

Figure 6: A scatter plot of the causal effect estimates between LHC-MR and IVW. Non-significant estimates lie on the axes, revealing a single trait pair that is significant for both methods but opposite in sign.

Given the apparent discrepancies between standard MR methods, we searched for trait-pairs whose causal effect estimates significantly differed between LHC-MR and at least two of the five standard MR methods. For eight such identified pairs we also applied a no-confounder-no-reverse causality model to closer mimic the assumptions of standard MR methods. Six of these discrepant pairs (BWeight-MI, HDL-LDL, LDL-SVstat, SBP-MI, SBP-SHeight, LDL-CAD) have not improved their agreement, hence the differences are due to the use of genome-wide markers (LHC-MR) instead of genome-wide significant ones alone (standard MR). Of note, for three of these pairs the best fitting LHC model had not detected any confounder. For the remaining two pairs (Edu-asthma, LDL-HDL), the substantially improved agreement upon using restricted LHC-MR suggests that the discrepancy was due to an ignored confounder. The increasing degree of agreement points to the possibility that standard MR methods yielded biased results due to the ignorance of a heritable confounder affecting these pairs.

To further support the existence of the confounders indicated by LHC-MR for 36 trait pairs, we used EpiGraphDB[28] to systematically identify those potential confounders. The database provided for each potential confounder a causal effect on trait X and Y ($r1$, and $r3$ in their notation), the ratio of which (r_3/r_1) was compared to our LHC-MR estimated t_y/t_x values representing the strength of the confounder acting on the two traits. For ten out of the 36 traits for which EpiGraphDB could identify reliable confounders, two are presented here (see Figure 7) and eight are shown in the supplement (Figs. S13). Notably, in all ten cases the average r_3/r_1 ratio agreed in sign with our t_y/t_x ratio, indicating that the characteristics of the confounders evidenced by LHC-MR agree with those found in an exhaustive confounder search. Since the reported causal effects in EpiGraphDB are not necessarily on the same (standardised)

scale as our LHC-MR estimates, we do not expect the magnitudes to agree. Figure 7 illustrates that the confounders of both the smoking-DM and SBP-DM relationships (evidenced by LHC-MR) may be obesity-related traits. Interestingly, for the LDL-HDL relationship, where the agreement between LHC- and standard MR estimates became more concordant upon enforcing a no-confounder-no-reverse causality LHC model, EpiGraphDB returned 28 potential confounders (Figure S13 panel a), whose causal effect ratios on these lipids (r_3/r_1) agreed both in sign and magnitude to our confounder effect ratio (t_y/t_x).

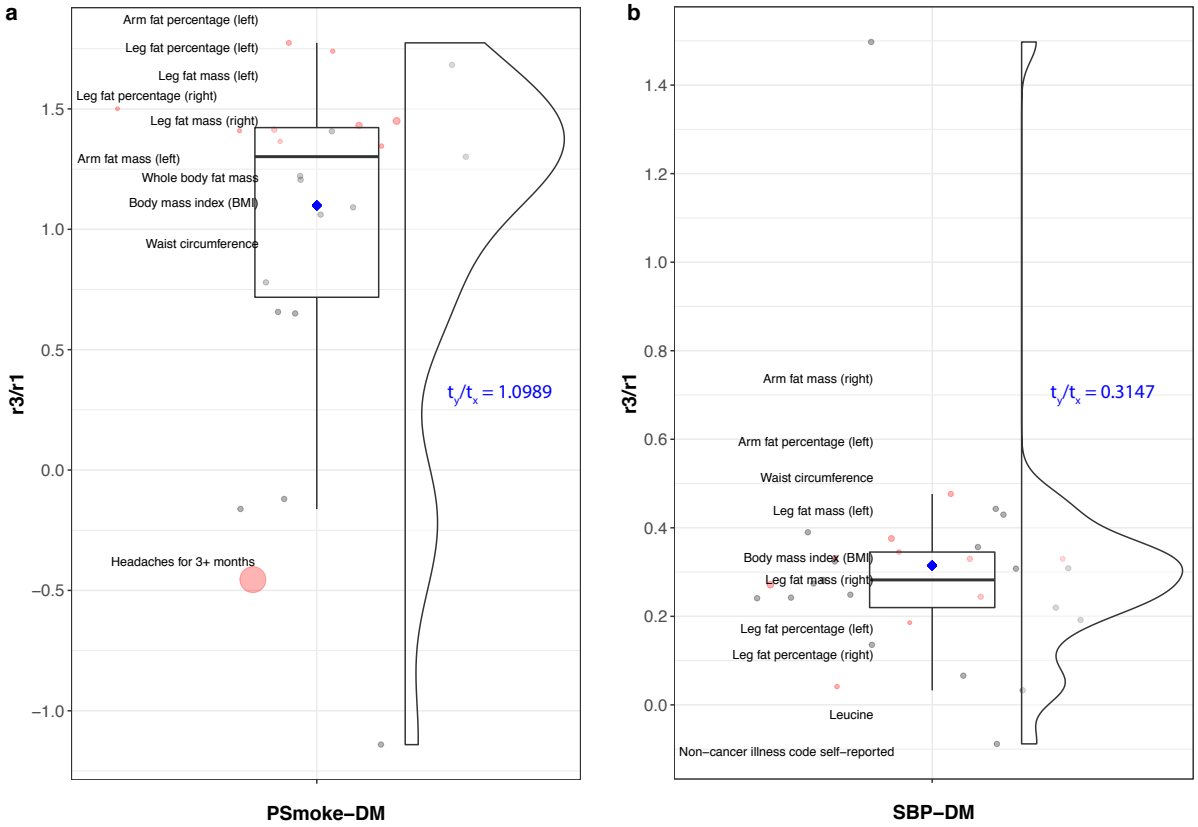

Figure 7: Confounders' effects obtained from EpiGraphDB plotted as a raincloud with the r_3/r_1 ratio on the y-axis. The blue diamonds represent the t_y/t_x ratio derived from the LHC model for that trait pair, also reported in blue text. Labelled dots in red and their varying size show the ten largest confounder traits in terms of their absolute effect product on the two traits, whereas grey dots represent the rest of the confounder traits found by EpigraphDB.

As described in the methods (Eq. 2), genetic correlation (r_G) can be computed from our estimated model parameters. Since our method and LD score regression[29] (LDSC) have slightly different assumptions (LHC-MR assumes a spike and slab effect size distribution (mixture of null and Gaussian) and a single heritable confounder), we first compared whether the two approaches produce similar genetic correlation estimates. We sampled 1,000 instances of the full parameter set of the LHC model (accounting for their estimated mean and covariance structure) and plugged them into Eq. 2 to yield 1,000 genetic correlation estimates for each trait pair. This facilitated the computation of the mean and variance of the estimated genetic correlations. As expected, we observe an overall good agreement between the two estimates (see Figure 8). While six pairs have opposite signs between genetic correlations estimated by LHC-MR and LDSC, only three of them are nominally significantly different (HDL-Asthma, HDL-SBP, SVstat-Asthma). Finally, our LHC-MR estimates were also used to estimate what fraction of the genetic covariance is due to bi-directional causal effects (by setting t_X and t_Y to zero in Eq. 2) and what is due to

heritable confounding. For example, we observed that none of the observed genetic correlation for SBP-DM, SVstat-DM are due to bi-directional causal effects, and it can mostly be put down to heritable confounding. On the contrary, Education-BMI and height-BMI genetic correlations are almost entirely driven by causal effects between them (see Table S3).

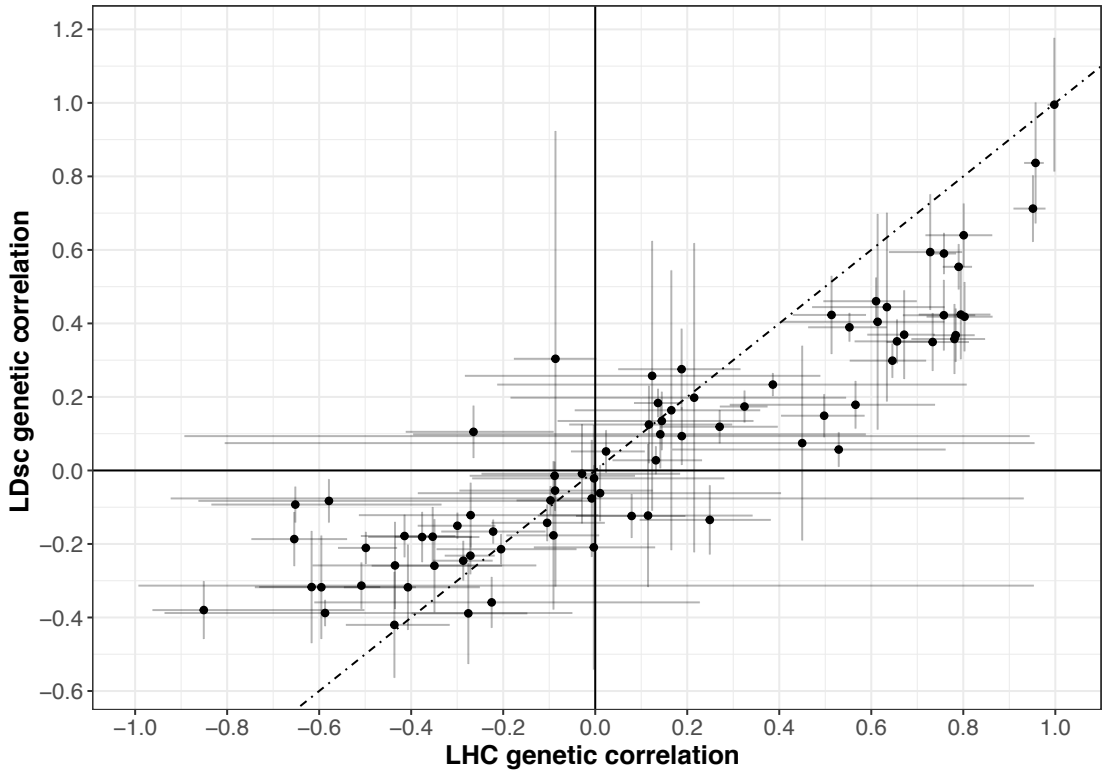

Figure 8: Scatter plot comparing the genetic correlation for each trait obtained from LDSC against the value calculated using derivations from the LHC-MR model, with a 95% confidence interval shown.

4 Discussion

We have developed a structural equation (mixed effect) model to account for a latent heritable confounder (U) of an exposure (X) - outcome (Y) relationship in order to estimate bi-directional causal effects between the two traits (X and Y). The method, termed LHC-MR, fits this model to association summary statistics of genome-wide genetic markers to estimate various global characteristics of these traits, including bi-directional causal effects, confounder effects, direct heritabilities, polygenicity, and population stratification.

We first demonstrated through simulations that in most scenarios, the method produces causal effect estimates with substantially less bias and variance than other MR tools. The direction and magnitude of the bias of classical MR approaches varied from scenario to scenario. Under standard settings with a small heritable confounder and no reverse causality present, all classical MR methods slightly underestimated the causal effect except for IVW. Interestingly, all standard MR methods showed a slight increasing trend in the estimate causal effect as the kurtosis of the underlying effect size distribution went up from 2 to 15. On the other hand, LHC-MR showed a barely noticeable decreasing trend with growing kurtosis. Polygenicity (π_X, π_U) did not increase the bias of any MR method, except for MR-Egger. However, as confounder causal effects (q_x, q_y) or its polygenicity (π_U) increased, all classical MR methods were prone to produce biased causal effect estimates with at least 10-fold higher variance than those of LHC-MR. Interestingly, mode-based estimators were robust to the presence of two concordant confounders, yet their bias is still 10-fold higher than LHC-MR. In summary, LHC-MR universally outperformed (in terms of bias, variance) all classical MR methods in virtually all tested scenarios, many of which violated its model assumptions.

Under two scenarios, our method lead to bimodal causal effect estimates (where one of the two estimates corresponded to the true underlying parameter value): when confounding is very severe (either being very heritable or having strong effect on X and/or Y) or when more than one confounder exist with opposite signed causal effects. In these scenarios, two equivalent models result with equal likelihood function values, but only one (heavily biased) is found by the classical MR methods. This visually translates to the situation where we cannot distinguish which slope represents the t_y/t_x ratio and which one corresponds to the causal effect $\alpha_{x\to y}$ in Figure S2.

We then applied our method to summary statistics of 13 complex traits from large studies, including the UK Biobank. We observed a general trend in our results that the liability to develop and to be diagnosed for a disease, e.g. diabetes, seems to improve (lower) BMI, while (in agreement with epidemiological studies) higher BMI is a risk factor for most diseases. We believe that this BMI-lowering effect of diabetes reflects a moderate lifestyle change upon the recommendation of medical personnel following the diagnosis.

Most robust MR methods (including MR Egger, median, mode and LHC) do not show a clear effect of smoking on diseases, which can be explained by the fact that most of smoking heritability is indirect ($t_x^2 = 0.053$), leaving very little room for direct heritability ($h_x^2 = 0.009$). This genetic architecture not only substantially reduces the power to detect outgoing causal effects, but may introduce bias in naive MR methods, such as IVW. For this reason, LHC-MR results are not reported in situations where direct heritability is smaller than 1%. We suspect that several heritable factors, such as education level and socio-economic status, heavily confound the smoking-disease relationship and may bias MR estimates. Note that we do not claim that smoking has negligible effect on diseases, but simply argue that MR is not the most ideal tool to reliably investigate its impact. Standard MR methods show no evidence for a causal effect of SBP on height, while our LHC-MR estimate is -0.3 ($P = 1.24 \times 10^{-70}$), which most probably reflects parental (maternal) effects as seen in previous studies[30]. This effect may be also be exacerbated by assortative mating that leads to an artificial correlation between the genetic

basis of this pair[31]. Similarly, IVW shows a strong negative effect of BMI on education (-0.17, $P = 1.83 \times 10^{-25}$), while LHC-MR shows no effect. Our null effect estimate is confirmed by standard MR analysis based on within-families effect estimates, which are much more protected against confounding such as socio-economic status [31].

Thirty six trait pairs showed a strong confounding effect, in the form of significant t_x and t_y estimates. These pairs were investigated for the presence of confounders using EpiGraphDB, and 10 of them returned possible confounders. Most prominent of these pairs was SBP-DM, where LHC-MR identifies a reverse causal effect and a confounder with effects $t_x = 0.38(P = 4.38 \times 10{-}48)$ and $t_y = 0.11(P = 3.03 \times 10{-}47)$ on SBP and DM respectively. For this pair, EpigraphDB confirmed several confounders related to body fat distribution and body shape (Figure 7 panel b). Note that EpiGraphDB causal estimates are not necessarily on the scale of SD outcome difference upon 1 SD exposure change scale, hence they are not directly comparable with the t_x/t_y ratio, but are rather indicative of the sign of the causal effect ratio of the confounder.

LHC-MR also enabled us to estimate the extent to which genetic covariance emerges due to bi-directional causal effects vs due to heritable confounding. For example, we evidence that genetic (and phenotypic) correlation observed for SBP-DM is mostly due to a confounder and is not driven by a causal relationship. On the other hand, the genetic correlation between education-BMI is driven by the education$\rightarrow$ BMI causal effect. The major difference between the genetic correlation obtained by LD score regression vs LHC-MR is that our model approximates all existing confounders by a single latent variable, which may be inaccurate when multiple ones exist with highly variable t_X/t_Y ratio. Our LHC-MR method is not designed to estimate genetic correlation, but it can be useful for qualitatively dissecting the contribution of causal effects and confounders to genetic correlation.

To our knowledge only two recent papers use similar models and genome-wide summary statistics. The LCV approach[32] is a special case of our model, where the causal effects are not included in the model, but they estimate the confounder effect mixed with the causal effects to estimate a quantity of genetic causality proportion (GCP). In agreement with others[10], we would not interpret non-zero GCP as evidence for causal effect. Moreover, in other simulation settings LCV has showed very low power to detect causal effects (by rejecting GCP=0) (Fig S15 in Howey et al.[33]). Another very recent approach, CAUSE[10], proposes a structural equation mixed effect model similar to ours. However, there are several differences between LHC and CAUSE: (a) we allow for bi-directional causal effects and model them simultaneously, while CAUSE is fitted twice for each direction of causal effect; (b) they use an adaptive shrinkage method to estimate β_x and β_y separately, while we fit all parameters at once; (c) CAUSE estimates the correlation parameter empirically; (d) we assume that direct effects come for a two-component Gaussian mixture, while they allow for larger number of components; (e) they account for LD by multiplying the observed (standardised) effects with the pairwise LD matrix, while we do not need to transform the observed effects, only introduce the LD score in the covariance matrix; (f) CAUSE adds a prior distribution for the causal/confounder effects and the proportion π_u, while LHC-MR does not; (g) to calculate the significance of the causal effect they estimate the difference in the expected log point-wise posterior density and its variance through importance sampling, whereas we can use a simple likelihood ratio test for nested models. Because of point (a), the CAUSE model is a special case of ours when there is no reverse causal effect. We have the advantage of fitting all parameters simultaneously, while they only approximate this procedure. Although they allow for more than two-component Gaussian mixture, for most traits with realistic sample sizes we do not have enough power to distinguish whether two or more components fit the data better. Therefore, we believe that a two component Gaussian is a reasonable simplification. Due to the more complicated approach described in points (e-g), CAUSE is computationally much more intense than LHC-MR, taking > 10 CPU-hours in contrast to our 1 CPU-minute run time. In terms of results, in simulations

we have seen that our approach remains robust (i.e. unbiased) to increased polygenicity (up to 100%) of the confounder U, while CAUSE has substantial false positive rate already when $\pi_u > 0.3$ in samples of size 40,000 (see Fig4a of Morrison et al.[10]).

Our approach has its own limitations, which we list below. Like any MR method, LHC-MR provides biased causal effect estimates if the input summary statistics are flawed (e.g. not corrected for complex population stratification, parental effects). As discussed above, when the confounding is severe or heterogeneous, two distinct models fit the data equally well which can result in bimodal parameter estimates. As opposed to classical MR methods that give a single (biased) causal effect estimate, ours detects the competing models, one of which agrees with the model used for the simulation. When the GWAS sample size is relatively low (<50,000), other MR methods can yield estimates with less variance (but more bias). Also, LHC-MR is not an optimal solution for traits whose genetic architecture is not accurately described by a two-component Gaussian mixture of effect sizes. This may be the case for the LDL-CAD trait pair, since LDL is influenced by a couple of very large effect SNPs and a three-component mixture would fit better. For such a case, where there is little polygenicity ($\pi_X < 1\%$) and this trait pair has low genetic correlation (-0.01), it might be wiser to use standard MR methods to estimate causality. In addition, trait pairs with multiple confounders with heterogeneous effect ratios can violate our single confounder approximation LHC model and can lead to biased causal effect estimates. Such important violations could be detected by discrepancy in genetic correlation estimates of LHC-MR *vs* LD score regression, but are out of the scope of the present paper.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 16389. LD scores were calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Z.K. was funded by the Swiss National Science Foundation (31003A_169929, 310030_189147). For computations we used the CHUV HPC cluster.
We would like to thank Jack Bowden, Valentin Rousson, Matthew Robinson, George Davey Smith, Thomas Richardson and Eleonora Porcu for their valuable feedback and comments on this manuscript.

Author contributions

Z.K. devised and directed the project. Z.K., N.M., and L.D. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Data availability

The origin of the summary statistics data used is referenced in Table S2. UK Biobank data can be downloaded from `http://www.nealelab.is/uk-biobank`. GLGC GWAS data can be downloaded from `http://csg.sph.umich.edu/willer/public/lipids2013/` and CAD GWAS data from `http://www.cardiogramplusc4d.org/data-downloads/`

Code availability

The source code for this work can be found on `github.com/LizaDarrous/LHC-MR`

References

[1] Fewell, Z., Davey Smith, G., and Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. American journal of epidemiology *166*, 646–655.

[2] Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., and Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. Nature reviews. Genetics *19*, 566–580.

[3] Barter, P. J., Caulfield, M., Eriksson, M., Grundy, S. M., Kastelein, J. J. P., Komajda, M., Lopez-Sendon, J., Mosca, L., Tardif, J.-C., Waters, D. D., et al. (2007). Effects of torcetrapib in patients at high risk for coronary events. The New England journal of medicine *357*, 2109–2122.

[4] Fordyce, C. B., Roe, M. T., Ahmad, T., Libby, P., Borer, J. S., Hiatt, W. R., Bristow, M. R., Packer, M., Wasserman, S. M., Braunstein, N., et al. (2015). Cardiovascular drug development: is it dead or just hibernating? Journal of the American College of Cardiology *65*, 1567–1582.

[5] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in medicine *27*, 1133–1163.

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

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

[8] Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genetic epidemiology *40*, 304–314.

[9] Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. International journal of epidemiology *46*, 1985–1998.

[10] Morrison, J., Knoblauch, N., Marcus, J., Stephens, M., and He, X. (2019). Mendelian randomization accounting for horizontal and correlated pleiotropic effects using genome-wide summary statistics. bioRxiv.

[11] Pearce, N. and Lawlor, D. A. (2016). Causal inference-so much more than statistics. International journal of epidemiology *45*, 1895–1903.

[12] Jordan, D. M., Verbanck, M., and Do, R. (2019). The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. bioRxiv.

[13] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics *47*, 291–295.

[14] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics *47*, 1236–1241.

[15] R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

[16] Neale Lab (2018). UK BioBank. http://www.nealelab.is/uk-biobank/.

[17] Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., et al. (2015). The uk10k project identifies rare variants in health and disease. Nature *526*, 82–90.

[18] Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS genetics *14*, e1007371.

[19] 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

[20] Clogg, C. C., Petkova, E., and Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. American Journal of Sociology *100*, 1261–1293.

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

[22] Allen, M., Poggiali, D., Whitaker, K., Marshall, T., and Kievit, R. (2019). Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. Wellcome Open Research *4*.

[23] Sun, D., Zhou, T., Heianza, Y., Li, X., Fan, M., Fonseca, V. A., and Qi, L. (2019). Type 2 diabetes and hypertension. Circulation research *124*, 930–937.

[24] Tomeo, C. A., Field, A. E., Berkey, C. S., Colditz, G. A., and Frazier, A. L. (1999). Weight concerns, weight control behaviors, and smoking initiation.

[25] Cawley, J., Markowitz, S., and Tauras, J. (2004). Lighting up and slimming down: the effects of body weight and cigarette prices on adolescent smoking initiation.

[26] Marouli, E., Del Greco, M. F., Astley, C. M., Yang, J., Ahmad, S., Berndt, S. I., Caulfield, M. J., Evangelou, E., McKnight, B., Medina-Gomez, C., et al. (2019). Mendelian randomisation analyses find pulmonary factors mediate the effect of height on coronary artery disease. Communications biology *2*, 119.

[27] Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., and Davey Smith, G. (2019). Within family Mendelian randomization studies. Human Molecular Genetics *28*, R170–R179.

[28] MRC IEU (2019). EpiGraphDB. http://epigraphdb.org/.

[29] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236–1241.

[30] Thomas, D., Strauss, J., and Henriques, M.-H. (1991). How does mother's education affect child height? The Journal of Human Resources *26*, 183–211.

[31] Brumpton, B., Sanderson, E., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., Havdahl, A., et al. (2019). Within-family studies for mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. bioRxiv.

[32] O'Connor, L. J. and Price, A. L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nature genetics *50*, 1728–1734.

[33] Howey, R., Shin, S.-Y., Relton, C., Smith, G. D., and Cordell, H. J. (2019). Bayesian network analysis incorporating genetic anchors complements conventional mendelian randomization approaches for exploratory analysis of causal relationships in complex data. bioRxiv.