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 heritabilities, 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 performed 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. Due to this limitation, additional evidence is required before developing public health interventions in a bid to reduce the future burden of 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.

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 and more advanced (such as CAUSE [10]) 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.

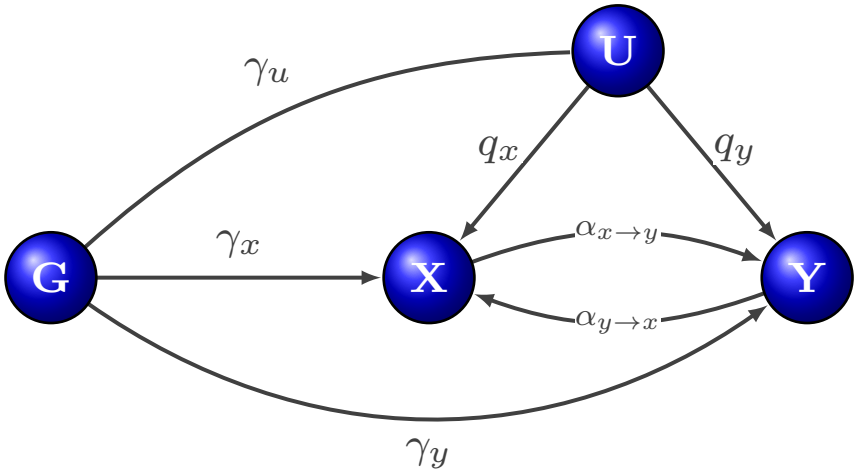


Figure 1: Schematic representation of the extended structural equation model (SEM). X and Y are two complex traits under scrutiny with a latent (heritable) confounder U with causal effects q_x and q_y on them. G represents a genetic instrument, with effects γ_x, γ_y and γ_u, respectively. Traits X and 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 on 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 $0 \leq \lambda_x, \lambda_y, \lambda_u \leq 1$ proportion of the genome have a direct effect on X, Y, U, respectively and these direct effects come from a Gaussian distribution. Namely,

$$\begin{aligned}
\boldsymbol{\gamma_x} &= \boldsymbol{\zeta_x} \odot \boldsymbol{\kappa_x} \quad \text{with} \quad \boldsymbol{\kappa_x} \sim \mathcal{N}(0, h_x^2/(m \cdot \lambda_x)) \quad \text{and} \quad \boldsymbol{\zeta_x} \sim \mathcal{B}(1, \lambda_x) \\
\boldsymbol{\gamma_y} &= \boldsymbol{\zeta_y} \odot \boldsymbol{\kappa_y} \quad \text{with} \quad \boldsymbol{\kappa_y} \sim \mathcal{N}(0, h_y^2/(m \cdot \lambda_y)) \quad \text{and} \quad \boldsymbol{\zeta_y} \sim \mathcal{B}(1, \lambda_y) \\
\boldsymbol{\gamma_u} &= \boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u} \quad \text{with} \quad \boldsymbol{\kappa_u} \sim \mathcal{N}(0, h_u^2/(m \cdot \lambda_u)) \quad \text{and} \quad \boldsymbol{\zeta_y} \sim \mathcal{B}(1, \lambda_u)
\end{aligned}$$

Here, $\odot$ denotes element-wise multiplication and $\mathcal{B}(1, q)$ the Bernoulli distribution. 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 $\lambda_x \cdot \lambda_y$, 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 j-th 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 $\boldsymbol{\rho_k} = \mathsf{G}_x' \cdot \boldsymbol{g}_k/n_x$ we get

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

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

$$\begin{aligned}
\widehat{\beta}_k^x &\approx (1 - \alpha_{y\to x} \cdot \alpha_{x\to y}) \cdot \widehat{\beta}_k^x = \boldsymbol{\rho_k'} \cdot ((\alpha_{y\to x} \cdot q_y + q_x) \cdot \boldsymbol{\gamma_u} + \alpha_{y\to x} \cdot \boldsymbol{\gamma_y} + \boldsymbol{\gamma_x}) + \eta_k^x \\
&= \boldsymbol{\rho_k'} \cdot ((\alpha_{y\to x} \cdot q_y + q_x) \cdot (\boldsymbol{\zeta_u} \odot \boldsymbol{\kappa_u}) + \alpha_{y\to x} \cdot (\boldsymbol{\zeta_y} \odot \boldsymbol{\kappa_y}) + (\boldsymbol{\zeta_x} \odot \boldsymbol{\kappa_x})) + \eta_k^x \\
&= (\alpha_{y\to x} \cdot q_y + q_x) \cdot (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_u'}) \cdot \boldsymbol{\kappa_u} + \alpha_{y\to x} \cdot (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_y'}) \cdot \boldsymbol{\kappa_y} + (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_x'}) \cdot \boldsymbol{\kappa_x} + \eta_k^x
\end{aligned}$$

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, assuming that the LD structures ($\boldsymbol{\rho_k}$) in the two samples are comparable, for $\widehat{\beta}_k^y$ estimated in the other sample (N_y) we obtain

$$\begin{aligned}
\widehat{\beta}_k^y &\approx \boldsymbol{\rho_k'} \cdot ((\alpha_{x\to y} \cdot q_x + q_y) \cdot \boldsymbol{\gamma_u} + \alpha_{x\to y} \cdot \boldsymbol{\gamma_x} + \boldsymbol{\gamma_y}) + \eta_k^y \\
&= (\alpha_{x\to y} \cdot q_x + q_y) \cdot (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_u'}) \cdot \boldsymbol{\kappa_u} + \alpha_{x\to y} \cdot (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_x'}) \cdot \boldsymbol{\kappa_x} + (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_y'}) \cdot \boldsymbol{\kappa_y} + \eta_k^y
\end{aligned}$$

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

2.3 Derivation of the likelihood function

Let π_x, π_y and π_u denote the fraction of the GWAS SNPs k for which $(\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_x}), (\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_y})$ and $(\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_u})$ is non-zero. These quantities represent the fraction of observed SNPs that are in non-zero LD with at least one causal variant for X, Y and U, respectively. As long as causal markers are sufficiently tagged by GWAS SNPs, $\pi_x \geq \lambda_x, \pi_y \geq \lambda_y$ and $\pi_u \geq \lambda_u$ because LD spreads association signals beyond strictly causal markers. We first work out the variances of the building blocks:

$$\begin{aligned}
\mathrm{Var}\left((\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_x'}) \cdot \boldsymbol{\kappa_x}\right) &= E\left[(\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_x})' \cdot \left(\frac{h_x^2}{m \cdot \lambda_x}\mathcal{I}\right) \cdot (\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_x})\right] = l_k \cdot \frac{h_x^2}{m} \\
\mathrm{Var}\left((\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_y'}) \cdot \boldsymbol{\kappa_y}\right) &= E\left[(\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_y})' \cdot \left(\frac{h_y^2}{m \cdot \lambda_y}\mathcal{I}\right) \cdot (\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_y})\right] = l_k \cdot \frac{h_y^2}{m} \\
\mathrm{Var}\left((\boldsymbol{\rho_k'} \odot \boldsymbol{\zeta_u'}) \cdot \boldsymbol{\kappa_u}\right) &= E\left[(\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_u})' \cdot \left(\frac{h_u^2}{m \cdot \lambda_u}\mathcal{I}\right) \cdot (\boldsymbol{\rho_k} \odot \boldsymbol{\zeta_y})\right] = l_k \cdot \frac{h_u^2}{m}
\end{aligned}$$

where $l_k = \boldsymbol{\rho_k'} \cdot \boldsymbol{\rho_k}$ as the LD score[13]. Since we know that the building blocks are spike-and-slab Gaussian mixtures, where the mixing proportions and total variances are known, we readily get the non-zero component variance as the overall variance divided by the mixing proportion of the

non-zero component. Thus the distributions of the three components can be written as follows

$$
\begin{aligned}
(\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_x}) \cdot \boldsymbol{\kappa_x} &\sim \pi_x \cdot \mathcal{N}\left(0, l_k \cdot \frac{h_x^2}{m \cdot \pi_x}\right) + (1-\pi_x) \cdot \mathcal{N}(0,0) \\
(\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_y}) \cdot \boldsymbol{\kappa_y} &\sim \pi_y \cdot \mathcal{N}\left(0, l_k \cdot \frac{h_y^2}{m \cdot \pi_y}\right) + (1-\pi_y) \cdot \mathcal{N}(0,0) \\
(\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_u}) \cdot \boldsymbol{\kappa_u} &\sim \pi_u \cdot \mathcal{N}\left(0, l_k \cdot \frac{h_u^2}{m \cdot \pi_u}\right) + (1-\pi_u) \cdot \mathcal{N}(0,0)
\end{aligned}
$$

Next, we partition the causal genomic variants into eight disjoint groups based on whether they are associated with (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, (0) 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 in LD with a causal variant effecting only U is $\pi_u \cdot (1-\pi_x) \cdot (1-\pi_y)$.

Therefore, the joint distribution of $(\widehat{\beta}_k^x, \widehat{\beta}_k^y)$ can be written as

$$
\begin{aligned}
\begin{pmatrix} \widehat{\beta}_k^x \\ \widehat{\beta}_k^y \end{pmatrix} &= \begin{pmatrix} \alpha_{y\to x} \cdot q_y + q_x \\ \alpha_{x\to y} \cdot q_x + q_y \end{pmatrix} \cdot (\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_u}) \cdot \boldsymbol{\kappa_u} + \begin{pmatrix} \alpha_{y\to x} \\ 1 \end{pmatrix} \cdot (\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_y}) \cdot \boldsymbol{\kappa_y} \\
&+ \begin{pmatrix} 1 \\ \alpha_{x\to y} \end{pmatrix} (\boldsymbol{\rho'_k} \odot \boldsymbol{\zeta'_x}) \cdot \boldsymbol{\kappa_x} + \begin{pmatrix} \eta_k^x \\ \eta_k^y \end{pmatrix} \\
&\sim \sum_{i=0}^{7} \omega_i \cdot \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_i\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\Omega_0 &= \Sigma_0, & \omega_0 &= (1-\pi_u)(1-\pi_x)(1-\pi_y) \\
\Omega_1 &= l_k \cdot \Sigma_u + \Omega_0, & \omega_1 &= \pi_u(1-\pi_x)(1-\pi_y) \\
\Omega_2 &= l_k \cdot \Sigma_x + \Omega_0, & \omega_2 &= \pi_x(1-\pi_u)(1-\pi_y) \\
\Omega_3 &= l_k \cdot \Sigma_y + \Omega_0, & \omega_3 &= \pi_y(1-\pi_u)(1-\pi_x) \\
\Omega_4 &= l_k \cdot (\Sigma_u + \Sigma_x) + \Omega_0, & \omega_4 &= \pi_u\pi_x(1-\pi_y) \\
\Omega_5 &= l_k \cdot (\Sigma_u + \Sigma_y) + \Omega_0, & \omega_5 &= \pi_u\pi_y(1-\pi_x) \\
\Omega_6 &= l_k \cdot (\Sigma_x + \Sigma_y) + \Omega_0, & \omega_6 &= \pi_x\pi_y(1-\pi_u) \\
\Omega_7 &= l_k \cdot (\Sigma_u + \Sigma_x + \Sigma_y) + \Omega_0, & \omega_7 &= \pi_u\pi_x\pi_y
\end{aligned}
$$

with

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

where $\rho_{x\cap y} := \frac{\rho_{x,y} \cdot n_{x\cap y}}{n_x \cdot n_y}$ is akin to 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 – just like with LD score regression – the trait correlation and sample overlap are not individually identifiable. To reduce the number of parameters we define $t_x := h_u \cdot q_x$ and $t_y := h_u \cdot q_y$ since h_u and q_x are separately not identifiable only their product. 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} is the correlation between GWAS SNPs k and j. The log-likelihood function is, thus, of the form

$$\log\left(\mathcal{L}\left(\boldsymbol{\theta}|\begin{pmatrix}\widehat{\boldsymbol{\beta}}^{\boldsymbol{x}}\\ \widehat{\boldsymbol{\beta}}^{\boldsymbol{y}}\end{pmatrix}\right)\right) \propto \sum_{k=1}^{K} w_k \cdot \left(\sum_{i=0}^{7} \omega_i \cdot \phi\left(\begin{pmatrix}\widehat{\beta}_k^x\\ \widehat{\beta}_k^y\end{pmatrix}; \begin{pmatrix}0\\0\end{pmatrix}, \Omega_i\right)\right) \tag{1}$$

Parameters $\{n_x, n_y, m, \boldsymbol{l}\}$ are known and the other 12 parameters

$$\boldsymbol{\theta} = \{i_x, i_y, \pi_x, \pi_y, \pi_u, h_x^2, h_y^2, t_x, t_y, \alpha_{x\to y}, \alpha_{y\to x}, \rho_{x\cap y}\}$$

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. Thus, we will have the following restrictions on the parameter ranges: 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 $\rho_{x\cap y} \in [-1/\max(n_x, n_y), 1/\max(n_x, n_y)]$.

2.4 Likelihood maximisation and standard error calculation

Our method, termed *Latent Heritable Confounder Mendelian Randomisation (LHC-MR)*, maximises this likelihood function to obtain the maximum likelihood estimate (MLE). 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 parameter estimates 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, under certain scenarios (related to high indirect heritability or very high polygenicity, exacerbated by lower sample size), we observed that two different sets of parameters can lead to an identical fit of the data resulting in two global optima. The reason for this is the difficulty in distinguishing the ratio of the confounder effects (q_y/q_x) from the causal effect ($\alpha_{x\to y}$), 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. Such bimodality can be observed at different levels: (i) For one given data generation using multiple starting points leads to different optima; (ii) For one given data generation different bootstrap samples result in parameter estimates that cluster around two distinct values; (iii) LHC-MR applied to multiple different data generations for a fixed parameter setting can yield different optima. All of these situations are signs of the same underlying phenomenon and most often co-occur.

By default we use 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 ($SE(\delta) = -|\widehat{\delta}|/\Phi^{-1}(P/2)$). To ensure that the standard errors are well-calibrated we also implemented the block jackknife procedure used by the LD score regression. For this we split the genome into 200 jackknife blocks and computed MLE in a leave-one-block-out fashion yielding $\widehat{\boldsymbol{\theta}}^{(-i)}, i = 1, \ldots, 200$ estimates. The standard error of the full SNP MLE was then defined as $Var(\widehat{\theta}) := \frac{m - m\cdot(1/200)}{m\cdot(1/200)} \cdot \frac{1}{200-1} \sum_{i=1}^{200} (\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^2 = \sum_{i=1}^{200} (\widehat{\boldsymbol{\theta}}^{(-i)} - \widehat{\boldsymbol{\theta}})^2$.

2.5 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 are then compared to those obtained by LD score regression.

2.6 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 ($\lambda_x = \pi_x = 0.1, \lambda_y = \pi_y = 0.15, \lambda_u = \pi_u = 0.05$), 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 the 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. We computed the false positive rate (for the causal effects or the confounders) of our method by simulating data with no causal effect, or with no confounder and then observed how LHC-MR estimates those parameters. 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. Second, 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. Third, we tested the assumption of the direct effects on our traits coming from a two-component Gaussian mixture by introducing a third component and observing how the estimates were effected. Simulations covered two scenarios, (1) introduced a large third component for X while decreasing the strength of U ($\pi_{x1} = 0.1, \pi_{x2} = 0.005, h^2_{x1} = 0.25, h^2_{x2} = 0.2, h^2_u = 0.1$), (2) introduced a large third component for both X and Y while decreasing U ($\pi_{x1} = 0.1, \pi_{x2} = 0.005, h^2_{x1} = 0.25, h^2_{x2} = 0.2, \pi_{y1} = 0.15, \pi_{y2} = 0.002, h^2_{y1} = 0.2, h^2_{y2} = 0.15, h^2_u = 0.1$). For each scenario, the data was simulated 100 times, and the likelihood function optimised to estimate the parameters. In order to explore whether removing SNPs with large effects could offer a simple way to eliminate the large effect component, we checked whether excluding SNPs with effects greater than certain arbitrary-set thresholds ($\beta^2 > 0.005/0.001/0.0005$) could improve parameter estimation.

2.7 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 disease 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 parameter estimates 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 3,781 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]).

2.8 Comparison against conventional MR methods and CAUSE

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 with effect sizes greater in the outcome than in the exposure (P-value < 0.05 in one-sided t-test). The default package settings for the clumping of SNPs ($r^2 = 0.001$) were used and the analysis was run with no further changes.

For real data, 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 only 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-value $< 0.05/(2*\text{number of comparison pairs})$), we used a restricted likelihood for LHC-MR allowing neither confounder nor reverse causal effect in order to mimic standard MR assumptions. The LHC-MR 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.

We also compared both our simulation as well as real data results against those of CAUSE[10]. We first generated simulated data under the LHC model following the standard settings and used it to estimate the causal effect using CAUSE. We then also generated simulated data using the CAUSE framework and used LHC-MR (as well as standard MR methods) to estimate the causal parameters. Lastly, we compared causal estimates obtained for the 156 trait pairs (78 bidirectional causal effects) from LHC-MR to those obtained when running CAUSE.

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 with X and Y with effects that are proportional (ratio $t_y : t_x$). 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, with 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

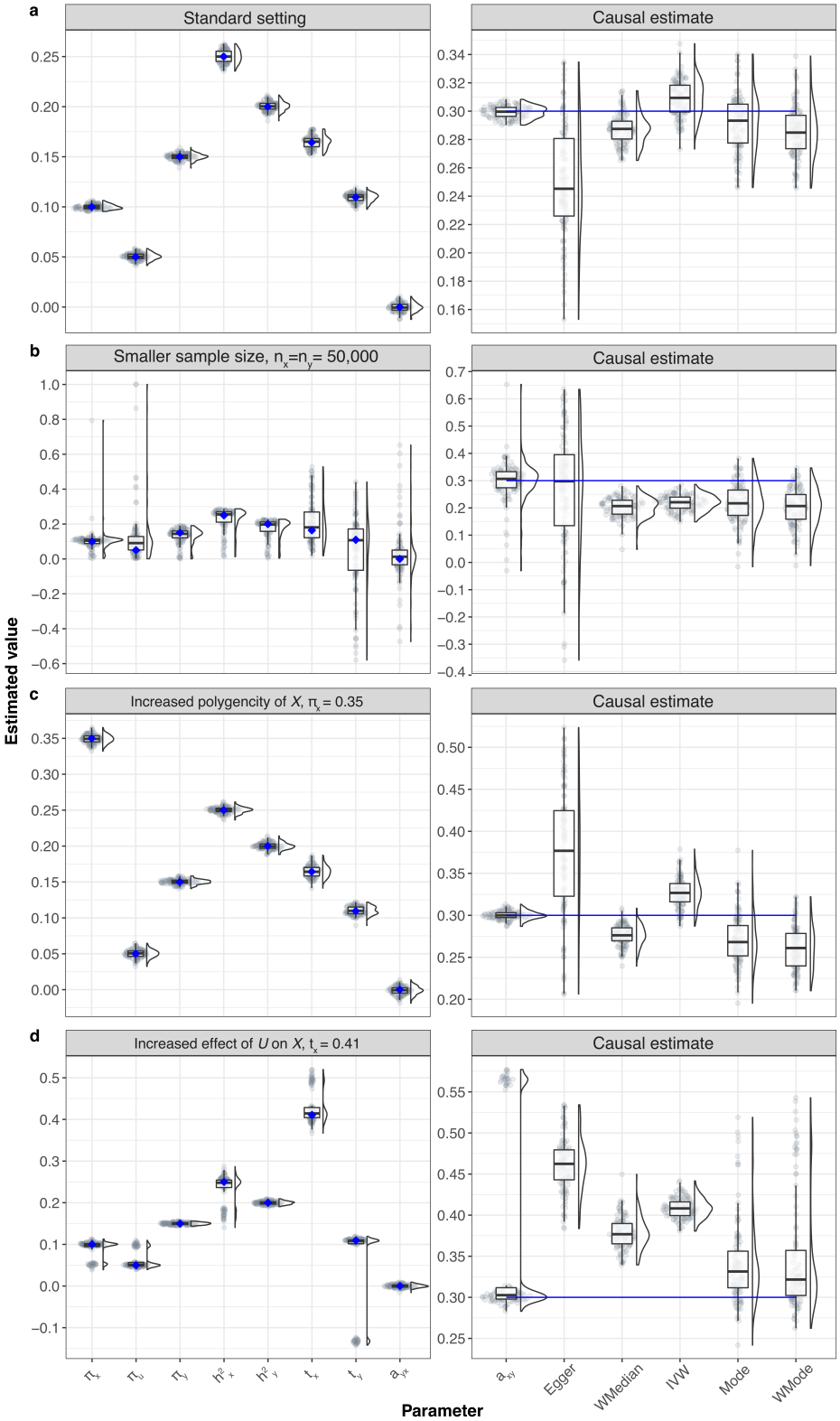


Figure 2: Raincloud boxplots[22] representing the distribution of parameter estimates from 100 different data generations under various conditions. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data 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.

biased (Figure S5) 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-MR 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 S6).

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 S7).

Reassuringly, our LHC-MR method is not influenced by the extent of polygenicity of X, Y or U either (Figure 2 panel *c*, Figure S8) 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 S9). 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 S10). In this case our method has the lowest bias and variance (bias = 5.4×10^{-4}, variance = 1.3×10^{-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 S8 panel *b*).

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 global maxima with identical likelihood function values (Figure 2 panel *d*). In such situations, the LHC-MR method proposes two equally likely underlying models with different causal effect estimates: $\widehat{\alpha}_{x\to y} = 0.57$ ($\mathrm{Var}(\widehat{\alpha}_{x\to y}) = 3.73 \times 10^{-5}$) and $\widehat{\alpha}_{x\to y} = 0.30$ ($\mathrm{Var}(\widehat{\alpha}_{x\to y}) = 3.27 \times 10^{-5}$)). 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 S11). Interestingly, other MR methods also exhibit kurtosis-dependent bias with a much more pronounced opposite trend. Second, we simulated summary statistics, where (contrary to our modelling assumptions) 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 = -1.3×10^{-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 S12). Conventional MR methods in this case almost all underestimated the causal effect with the exception of MR-Egger that had the lowest bias (bias = 8.4×10^{-3}). Third, we simulated effect sizes coming from a three-component Gaussian mixture distribution (null/small/large effects), instead of the classical spike-and-slab assumption of our model. Results revealed that the presence of a third component biases parameter estimation only if the introduced third component has a very strong and different effect than the second component, and if the confounder has a moderate effect. Such situations can lead to the selection of a no-confounder model, even when a true U exists (Supplementary Figure S13 panel *a*), and slightly bias the causal effect estimates by either merging true effect estimates with confounder effect or

leading to a significant estimate of b when no reverse causal effect estimate has been simulated (Supplementary Figure S13 panel a). In particular, for the first scenario, we observed bimodal parameter estimates for pi_x, pi_u and h_x^2 (one peak corresponded to the true value), a large overestimation of t_x (third- and possibly part of the second- component acting through U), and t_y estimated as null (non-significant). Reverse causal effect estimate was slightly overestimated and significant (median p-value $= 8.93 \times 10{-}04$). However, filtering SNPs with strong effects based on a threshold (0.0005 in Figure S13 panel b) before running LHC-MR allowed us to eliminate the third component and get back correct estimates. However, the optimal effect size threshold depends on the simulation parameters, hence its choice is not straightforward. Note that the underestimation of h_x^2 is entirely expected after filtering out SNPs strongly associated with X.

In the basic simulation setting we examined the coverage of the 95% confidence interval (CI) computed using SEs derived from both (block) jackknife and LRT. True parameter values for $\alpha_{x\to y}$, t_x and t_y fell into their 95% CI 95, 91 and 96 times out of 100 times respectively. The coverage of the LRT-based SE estimation for the same parameters revealed an inclusion of 100, 94 and 100 times out of 100 times. Thus we can conclude that while block jackknife-based SEs are accurate, its computation is 200-times slower than those of LRT-based SEs, which on the other hand are slightly more conservative. For real data application, the ratio of the SE obtained from the LRT to the block jackknife SE of several parameters from several trait pairs ranged from 0.26 to 1.21 with a median of 0.71, with the exception of the ratio for the t_x parameter (0.0039) in the BMI-DM relationship (see Supplementary Table 5). The observed discrepancy between the simulation and real data results is due to situations with bimodal likelihood surface, which occurred more often in our real data application. In such cases, the 95% confidence interval should be two distinct intervals around the two maxima, which is captured by LRT (thanks to using multiple starting points), but mistakenly inflated by the block jackknife SE calculation, which is overly conservative in these scenarios. As a consequence, block jackknife-based SEs would lead to incorrect conclusions: even if two (non-zero) parameter values are equally probable, we could still confidently reject the null hypothesis. For these reasons, we chose to use LRT-based SEs for real data applications. Note that these situations do not occur for LD score regression, where such approach for SE estimation is correct.

When running CAUSE on data simulated using the LHC-MR framework in order to estimate a causal effect (γ), we investigated three different scenarios, each with multiple data generations: one where the underlying model has a causal effect of 0.3 only, another where the underlying model has a shared factor/confounder with effect on both exposure and outcome, a third where the underlying model has both a causal effect and a shared factor. The data generated using the LHC-MR model was done under the standard settings ($\pi_x = 0.1, \pi_y = 0.15, \pi_u = 0.05$, $h_x^2 = 0.25, h_y^2 = 0.2, h_u^2 = 0.3$, $t_x = 0.16, t_y = 0.11$, $\alpha_{x\to y} = 0.3, \alpha_{y\to x} = 0$, $m = 50k, n_x = n_y = 500k$). For each setting 100 different replications were investigated.

When there was an underlying causal effect only, CAUSE preferred the causal model only 3% of the times. It slightly over estimated γ to 0.313, whereas the estimation of the other parameters η and q were small (around 0 and 0.05 respectively), as seen in Figure S14. In the case of an underlying shared effect only, CAUSE preferred the shared model 100% of the time, and thus there was no causal estimation. In the third case, and in the presence of both, CAUSE preferred the sharing model in 96% of the simulations.

In the reversed situation, where data was generated using the CAUSE framework (with parameters $h_1 = h_2 = 0.25, m = 50k, N1 = N2 = 500k$) and LHC-MR was used to estimate the causal effect, we saw the following results (see Figure S15). First, when we generated data in the absence of causal effect ($\gamma = 0, \eta = sqrt(0.05), q = 0.1$), CAUSE does extremely well in estimating a null causal effect, whereas LHC-MR identifies the correct null value 69% of the

simulations, but also provides another estimate 0.225 (in 31% of the simulations). Second, in the absence of a confounder ($\gamma = 0.3, \eta = 0, q = 0$), CAUSE underestimates the causal effect substantially ($\widehat{\gamma} = 0.24$) compared to LHC-MR which accurately estimates the causal effect to an average of 0.30 over the 100 runs. Finally, in the presence of both a confounder and a causal effect ($\gamma = 0.3, \eta = sqrt(0.05), q = 0.1$), CAUSE also underestimates the causal effect substantially ($\widehat{\gamma} = 0.23$), but LHC-MR provides a bimodal estimate ($\widehat{\gamma} = 0.30$ and 0.08 (58%,42% of the times, respectively). Interestingly, classical MR methods also outperform CAUSE.

3.3 Application to association summary statistics of complex traits

We applied our LHC-MR and other MR methods to estimate all pairwise causal effects between 13 complex traits. These results are presented as a heatmap in Figure 3 and as a table in Supplementary Table 4. Our comparison of the results obtained by LHC-MR and standard MR methods is detailed below and represented in Figure S16. In summary, for 153 out of 156 trait pairs, causal estimates were not significantly discordant between the five different standard MR methods. Thirty-five out of 153 pairs were not suitable for comparison either due to low estimated direct heritability from LHC-MR (< 0.01) (i.e. not enough instrument strength) or a bimodal estimated causal effect (i.e. different likelihood maximisation starting points led to distinct parameters with identical likelihood function value). For 117 out of those 118 comparable trait pairs, our LHC-MR causal effect estimates were concordant (not significantly different) with at least two out of five standard MR methods' estimates. Only one pair disagreed with four MR methods, three other pairs were discordant with three MR methods, and two pairs had conflicting estimates between LHC-MR and two 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 (e.g. coronary artery disease (CAD) and obesity). Our results reveal that many diseases lead to 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.

Interestingly, causal effects of height on SBP and DM have been previously seen 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 $0.52 (P = 2.38 \times 10^{-289})$ and $-0.05 (P = 4.41 \times 10^{-53})$ 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) or may reflect parental/family effects. Indeed, 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 20% ($P = 4.57 \times 10^{-30}$).

A systematic comparison between IVW and LHC-MR has shown generally good agreement between the two methods, which is illustrated in Figure 4. 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

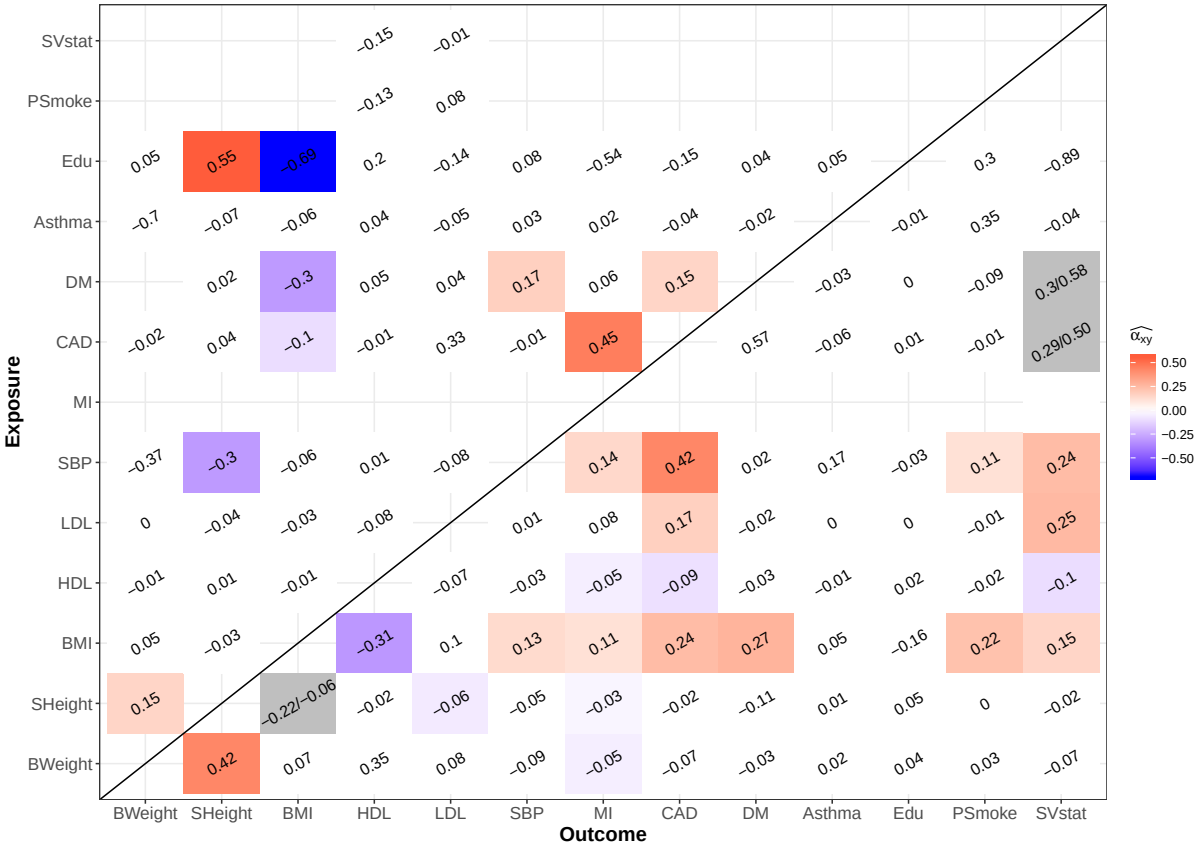


Figure 3: Heatmap representing the bidirectional causal relationship between the 13 UK Biobank traits from the winning most nested (parsimonious) models. The causal effect estimates in coloured tiles all have a significant p-values surviving Bonferroni multiple testing correction with a threshold of 3.2×10^{-4}. When two distinct (global) maxima exist we report the means of each peak in that distribution in the grey boxes (detailed in Figures S17 and S18). 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 causal 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

estimate, or significant with a disagreeing sign. The diagonal (seen in Figure 4) representing the agreement in significance status and sign between the two methods, is heavily populated. On the other hand, seven 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 17 pairs. We believe that many of these 17 pairs may be false positives, since only five of them 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 S21.

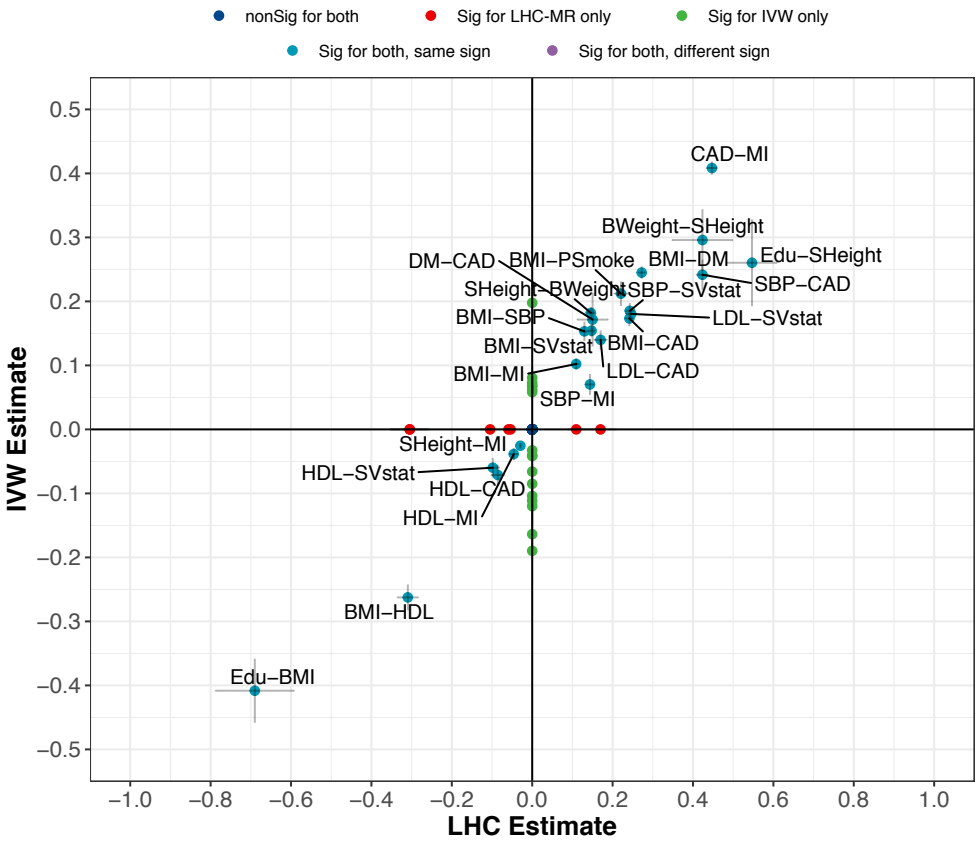


Figure 4: A scatter plot of the causal effect estimates between LHC-MR and IVW. Non-significant estimates lie on the axes, while significant estimates for both methods are seen in the diagonal.

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 (see Supplementary Table 7). For three (SBP-CAD, LDL-SVstat, SBP-MI) of the six discrepant pairs the winning model had no confounder or reverse causation and only one (BMI-CAD) of the other three pairs (Edu-BMI, Edu-SHeight, BMI-CAD) improved its agreement with standard MR methods upon restricting the LHC model to a no-confounder-no-reverse causality scenario (mimicing the assumptions of standard MR methods). Hence we suspect that the differences are due to the use of genome-wide markers (LHC-MR) instead of genome-wide significant ones alone (standard MR).

To support the existence of the confounders indicated by LHC-MR for 38 trait pairs, we used EpiGraphDB[28, 29] 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. EpiGraphDB could identify reliable confounders for thirteen out of the 38 traits. Two are presented here (see Figure 5) and

the rest are shown in the supplement (Figs. S20). Notably, in 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 (including all traits in the UK Biobank). 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 5 illustrates that the confounders of both the SBP-Edu and SBP-DM relationships (evidenced by LHC-MR) may be obesity-related traits.

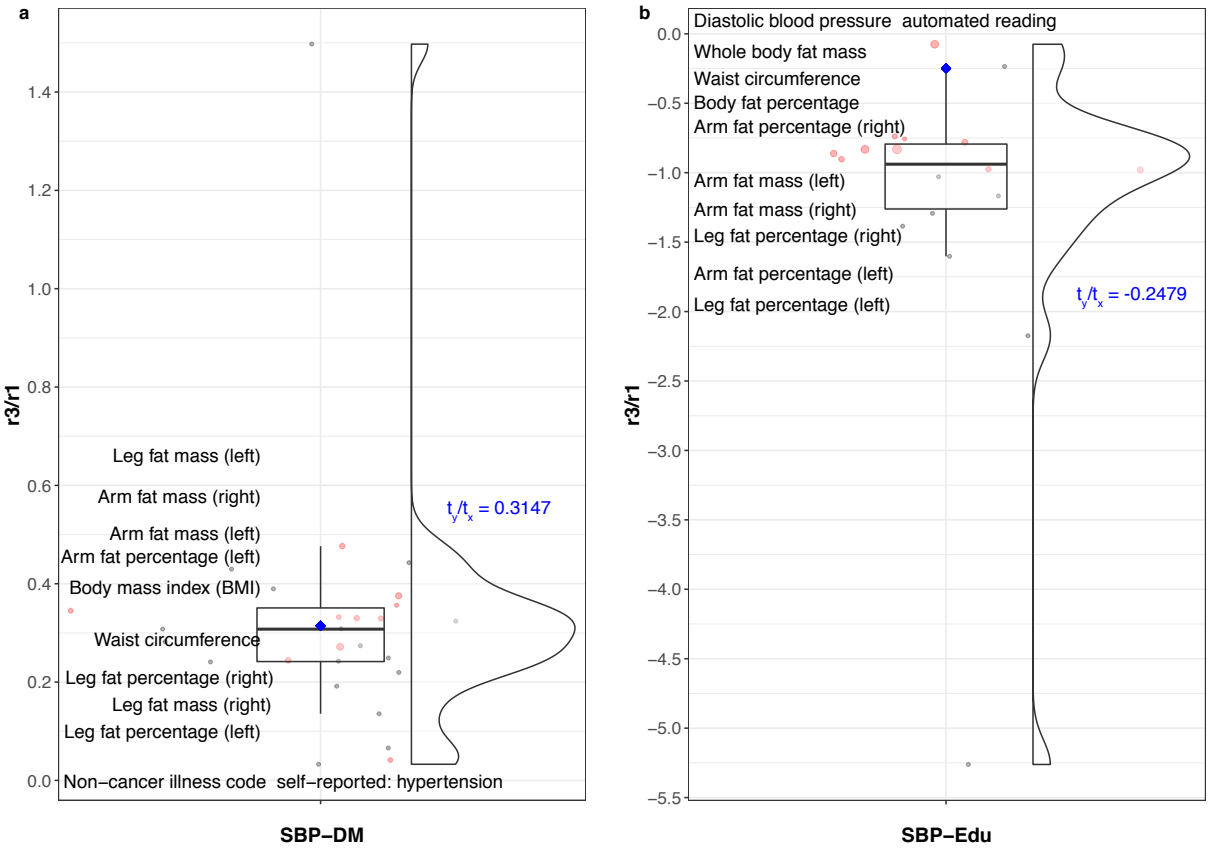


Figure 5: Confounders' effects obtained from EpiGraphDB plotted as a jitter plot with the r_3/r_1 ratio on the y-axis. The blue diamonds represent the t_y/t_x ratio derived from the LHC model for that trait pair, also reported in blue text. Labelled dots in red and their varying size show the ten largest 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[30] (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 6). 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). 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 attributed to heritable confounding. On the contrary, the genetic correlation of the pairs Edu-BMI ($\alpha_{x\to y} = -0.69$, $P = 2.07 \times 10^{-12}$) and height-BMI ($\alpha_{x\to y} = -0.23$, $P = 1.40 \times 10^{-10}$) are almost entirely driven by causal effects between them (see Table S3).

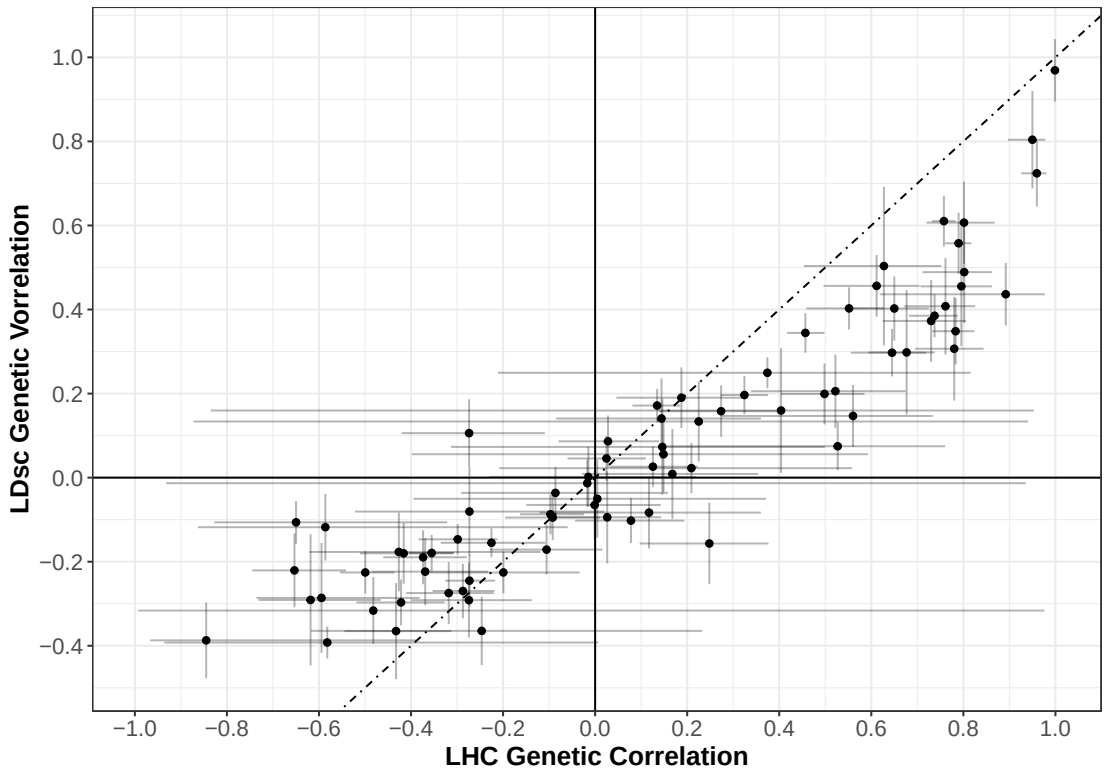


Figure 6: Scatter plot comparing the genetic correlation for each trait obtained from LDSC against the value calculated using derivations from the LHC-MR model, with a 95% confidence interval shown.

As for the comparison of LHC-MR against CAUSE for real trait pairs, we ran CAUSE on all 156 trait pairs (bi-directional), and extracted the parameter estimates that corresponded to the methods winning model. The p-value threshold was corrected for multiple testing and was equivalent to 0.05/156. Based on that threshold, the p-value that compared between the causal and the sharing model of CAUSE was used to choose one of the two. Then the parameters estimated from the winning model, γ (only for causal model), η and q, were compared to their counterparts in LHC.

Whenever the causal effect estimates when significant both for CAUSE (γ) and LHC-MR, they always agreed in sign (Table 1) with a moderate Spearman correlation of 0.556. When compared to the causal effect estimate from IVW, LHC-MR was highly correlated (0.949), whereas CAUSE had only a moderate correlation (0.507).

Similarly, the confounder effect ratio of LHC-MR ($t_y : t_x$) can be compared to the confounder effect estimate of CAUSE (η). These confounding quantities by CAUSE and LHC-MR agreed in sign for all but one trait pair (with the exception of LDL↔HDL), with a Spearman correlation of 0.739.

Finally, we explored whether significantly non-zero MR-Egger intercept could be indicative of a latent heritable confounder. Namely, out of 156 trait pairs, MR-Egger returned 21 nominally significant intercepts ($P < 0.05$). Of those 21, LHC-MR had 10 trait pairs with significant

Pair	$\alpha_{x \to y}$	**p-value**	γ	**IVW** $\alpha_{x \to y}$	**p-value**
BMI-CAD	0.2418	1.41E-126	0.28 (0.25, 0.32)	0.1733	7.6E-48
BMI-DM	0.2724	1.51E-201	0.04 (0.04, 0.05)	0.2450	1.52E-142
BMI-HDL	-0.3094	2.70E-32	-0.15 (-0.17, -0.12)	-0.2625	2.63E-38
BMI-MI	0.1094	6.81E-45	0.01 (0.01, 0.01)	0.1022	1.66E-31
BMI-PSmoke	0.2208	1.75E-38	0.16 (0.13, 0.18)	0.2121	2.44E-28
BMI-SBP	0.1299	2.02E-29	0.13 (0.11, 0.14)	0.1532	4.68E-24
BMI-SVstat	0.1478	3.30E-70	0.03 (0.03, 0.03)	0.1542	4.56E-59
BWeight-SHeight	0.4235	2.42E-08	0.34 (0.3, 0.39)	0.2959	7.94E-10
CAD-MI	0.4471	2.41E-233	0.02 (0.02, 0.02)	0.4084	0
CAD-SVstat	0.29/0.50	2.76E-29	0.03 (0.03, 0.04)	0.3201	2.49E-34
DM-SVstat	0.3/0.58	6.89E-16	0.4 (0.34, 0.46)	0.2530	1.25E-16
Edu-BMI	-0.6901	2.07E-12	-0.34 (-0.4, -0.28)	-0.4083	2.75E-16
Edu-SHeight	0.5468	4.14E-18	0.17 (0.13, 0.22)	0.2604	0.00013
LDL-CAD	0.1703	2.76E-25	0.36 (0.28, 0.43)	0.1400	3.67E-21
LDL-SVstat	0.2464	1.04E-72	0.06 (0.05, 0.06)	0.1804	4.65E-107
SBP-CAD	0.4234	1.70E-159	0.45 (0.39, 0.51)	0.2416	1.14E-24
SBP-SHeight	-0.3047	7.03E-10	-0.13 (-0.15, -0.1)	-0.0947	0.00552
SBP-SVstat	0.2427	6.92E-103	0.04 (0.04, 0.05)	0.1852	3.32E-52
SHeight-BMI	-0.22/-0.06	1.40E-10	-0.13 (-0.14, -0.11)	-0.0842	3.93E-22
SHeight-BWeight	0.1459	3.23E-49	0.24 (0.22, 0.25)	0.1820	4.79E-90

Table 1: Table comparing the causal estimates of LHC-MR, CAUSE, and IVW for trait pairs that had a significant causal effect in all three methods. The column showing the gamma (causal effect) estimate of the CAUSE method also reports its 95% credible intervals.

confounder effect. For nine out of these 10 pairs (with the exception of BMI-SHeight), the $t_y : t_x$ confounder ratio of LHC-MR agreed in sign with the MR-Egger intercept, indicating that often directional pleiotropy may reflect a heritable confounder.

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, polygenicities, 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_y) 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 sign 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 little 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[31,32]. This effect may also be exacerbated by assortative mating that leads to an artificial correlation between the genetic basis of this pair[33]. Similarly, IVW shows a strong negative effect of BMI on education (-0.20, $P = 4.57 \times 10^{-30}$), 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 [33].

Thirty eight 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 13 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 $t_x = 0.34(P = 4.38 \times 10^{-48})$ 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 5 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 found 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 Edu-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[34] 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[10,35], 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.[36]). Another very recent approach, CAUSE[10], proposes a structural equation mixed effect model similar to ours. However, there are several differences between LHC-MR 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 from 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 (or block jackknife). 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 up to 1.25 CPU-hours in contrast to our 1 CPU-minute run time for a single starting point optimisation (which is massively parallelisable). 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]). When we compared the performance of CAUSE and LHC-MR both on data simulated based on the respective models, we found that while LHC-MR yielded accurate causal effect estimates for both kinds of underlying models, CAUSE exhibits more biased effect estimates and a general bias towards the confounding model (conservative). It is understandable since the primary aim of CAUSE is model selection and it is slightly less geared towards parameter estimation, especially for settings where both sharing and causal effects are present leading to very broad estimates. Also, CAUSE parameter estimates were sometimes sensitive to the choice of the prior.Furthermore, we noted that summary statistics simulated based on the CAUSE model more often lead to bimodal causal effect estimates, i.e. the parameters of such a model seem to be less identifiable both for CAUSE and LHC-MR. Finally, when applying both LHC-MR and CAUSE to 156 complex trait pairs, we observed that the causal effects are reasonably well correlated and agreed in sign for trait pairs deemed significantly causal by both methods. In addition, LHC-MR causal estimates were more similar to those of IVW than the estimates provided by CAUSE.

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 may fit better. Also, when there is too little polygenicity ($\pi_X < 1\%$) and/or low direct heritability ($< 1\%$), 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 the single confounder LHC model and can lead to biased (or bimodal) 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 mathematical derivations, 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 summary statistics 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. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature Genetics *52*, 740–747.

[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] Liu, Y., Elsworth, B., Erola, P., Haberland, V., Hemani, G., Lyon, M., Zheng, J., and Gaunt, T. R. (2020). Epigraphdb: A database and data mining platform for health data science. bioRxiv.

[30] 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.

[31] 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.

[32] Warrington, N. M., Beaumont, R. N., Horikoshi, M., Day, F. R., Helgeland, Ø., Laurin, C., Bacelis, J., Peng, S., Hao, K., Feenstra, B., et al. (2019). Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet *51*, 804–814.

[33] 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.

[34] 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.

[35] Brown, B. C. and Knowles, D. A. (2020). Phenome-scale causal network discovery with bidirectional mediated mendelian randomization. bioRxiv.

[36] 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.

Supplementary Materials & Methods

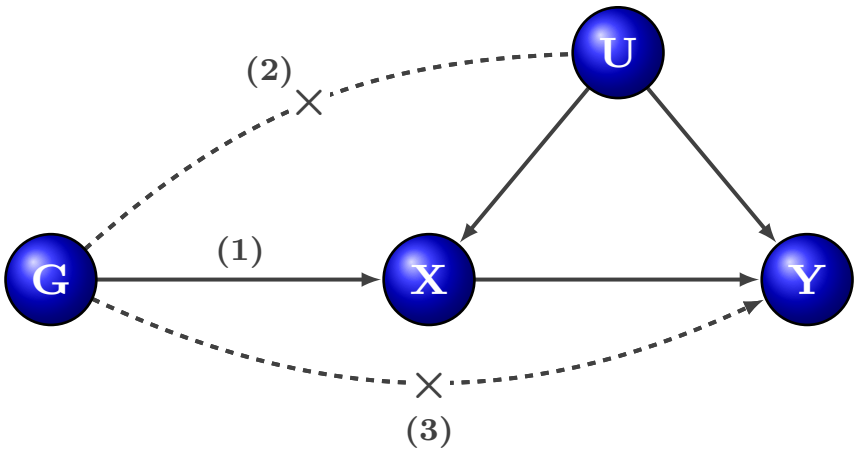


Figure S1: Basic assumptions of Mendelian randomisation. (1) Relevance – genetic data, denoted by G, is robustly associated with the exposure. (2) Exchangeability – G is not associated with any confounder of the exposure-outcome relationship. (3) Exclusion restriction – G is independent of the outcome 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).

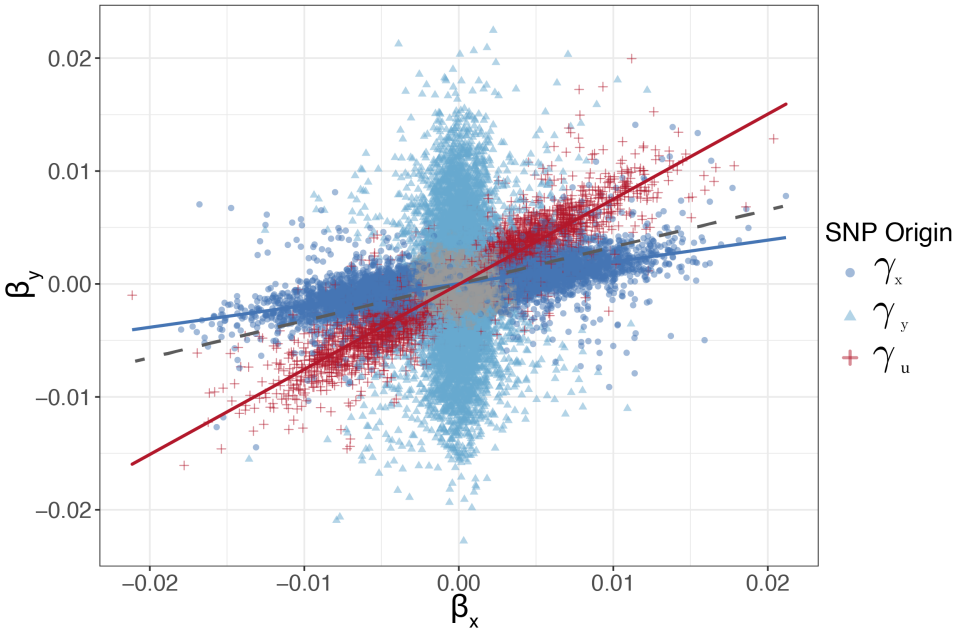


Figure S2: A plot of simulated observed SNP effects on traits X and Y, coloured by the strongest effect between the three vectors $\gamma_x, \gamma_y, \gamma_u$. SNPs in grey are those with no effect on any of the traits. This illustration shows the distinct clusters that could arise in the presence of a confounder. The dark blue cluster of SNPs represents those that are not in violation of any of the MR assumption, and hence its slope reflects the true causal effect of X on Y, while the red cluster of SNPs are those associated with the confounder. The steeper slope of the red cluster of SNPs causes a typical regression line - shown in grey - that represents the causal effect (estimated using conventional MR methods) to be overestimated.

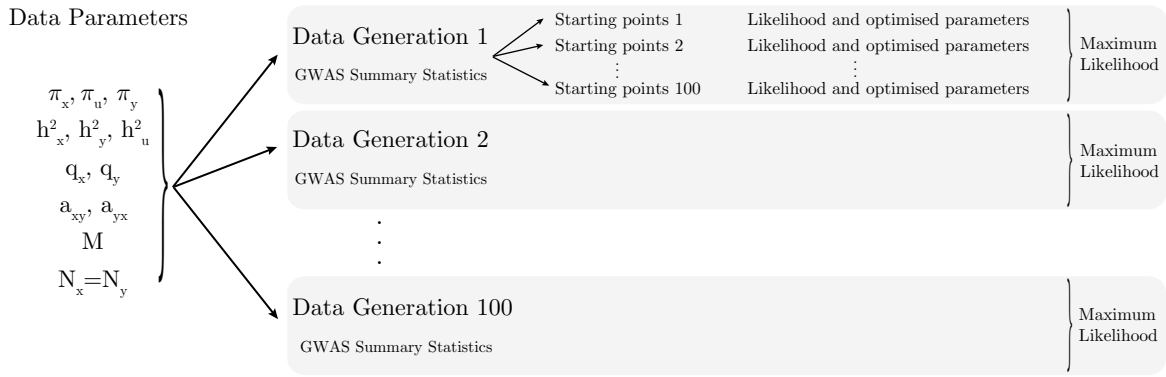


Figure S3: A schema showing the workflow of the simulation results. For a single set of parameter settings, 100 different data generations of GWAS summary statistics are created for trait X and Y. The summary statistics of a single data generation, as well as the sample size and SNP number, are used in the likelihood optimisation function that is run with 100 different random starting points in order to explore the likelihood surface. A single maximum likelihood and its corresponding estimated parameters are selected to represent the estimates of that data generation. And this is repeated for the other generations. The results for several data generation are often represented in boxplots throughout the paper.

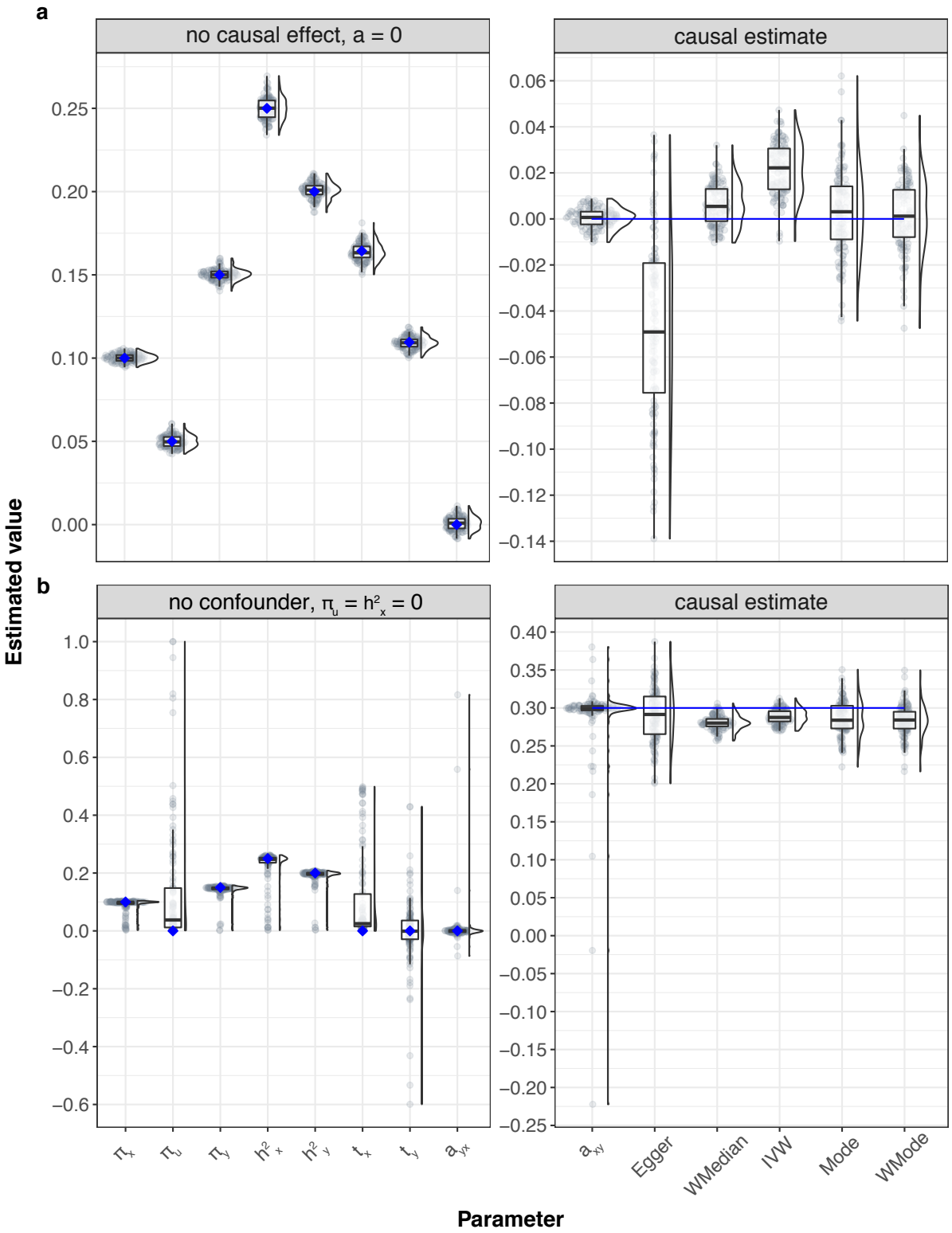


Figure S4: Raincloud boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. In panel *a*, the data simulated had no causal effect in either direction. A false positive test was run by comparing the likelihoods of the complete model and the nested model that removes the casual effect $\alpha_{x\to y}$. In 95 times out of 100, the LRT test favoured the nested model over the complete one. In panel *b*, the data simulated had no confounder effect with π_u, t_x, and $t_y = 0$. Similarly, in 93 times out of 100, the nested model was favoured by the LRT over the complete model.

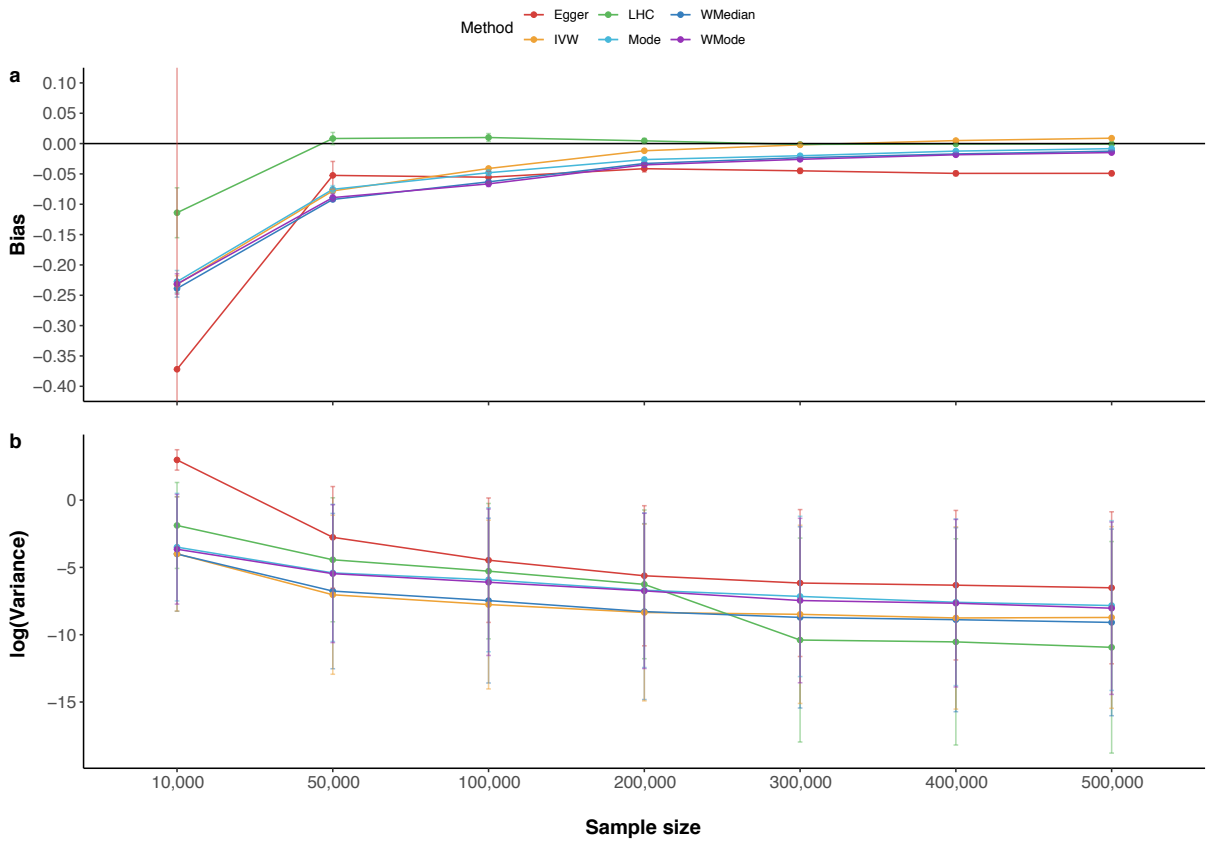


Figure S5: Line plots representing the bias and log variance of the estimated causal parameter, $\alpha_{x \to y}$, using the various MR methods against LHC-MR for different sample size ranging from 10,000 to 500,000 with the standard error reported.

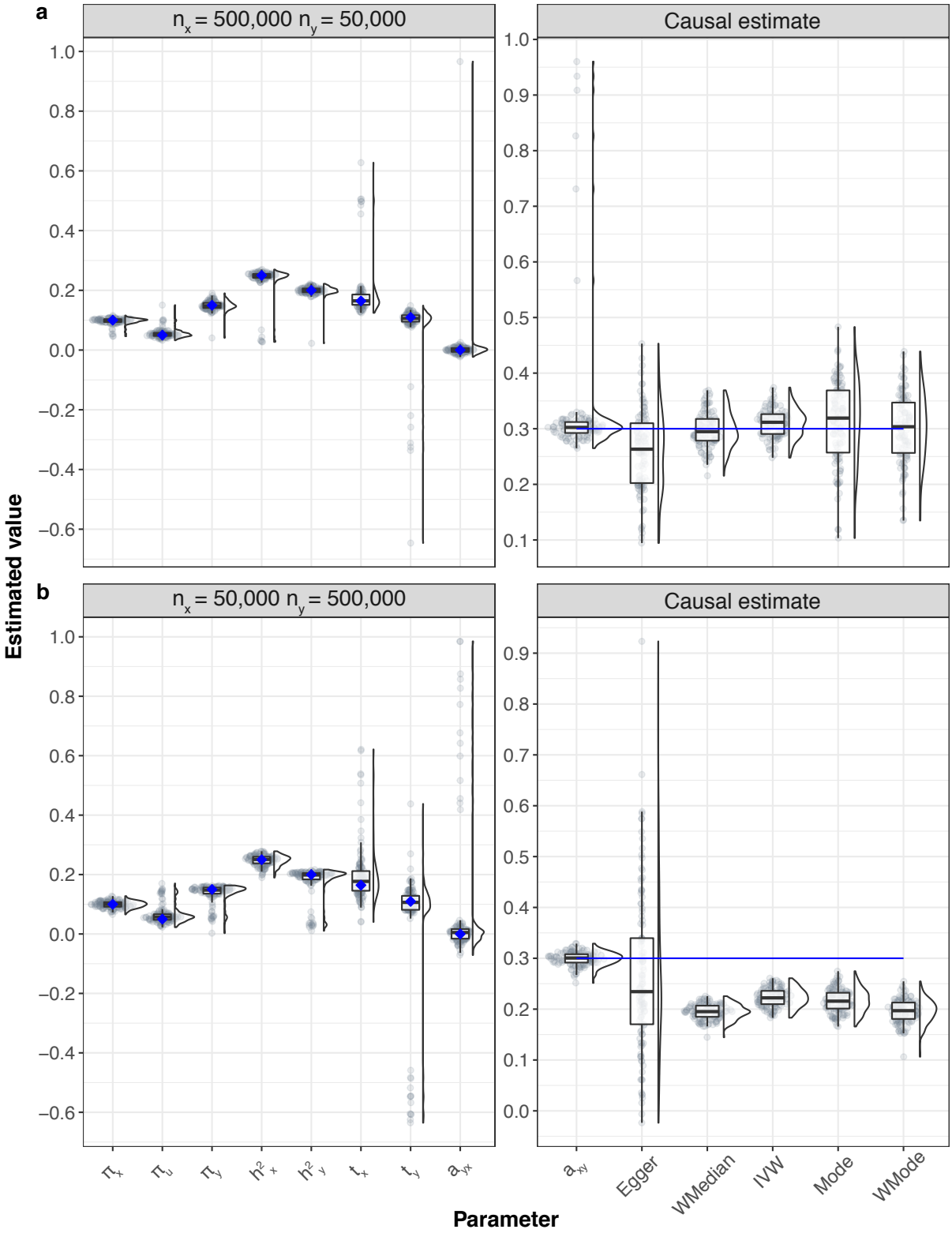


Figure S6: Raincloud boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. In this figure, samples sizes for the two traits differ as such n_x = 500,000 and n_y = 50,000 for panel *a*, and n_x = 50,000 and n_y = 500,000 for panel *b*.

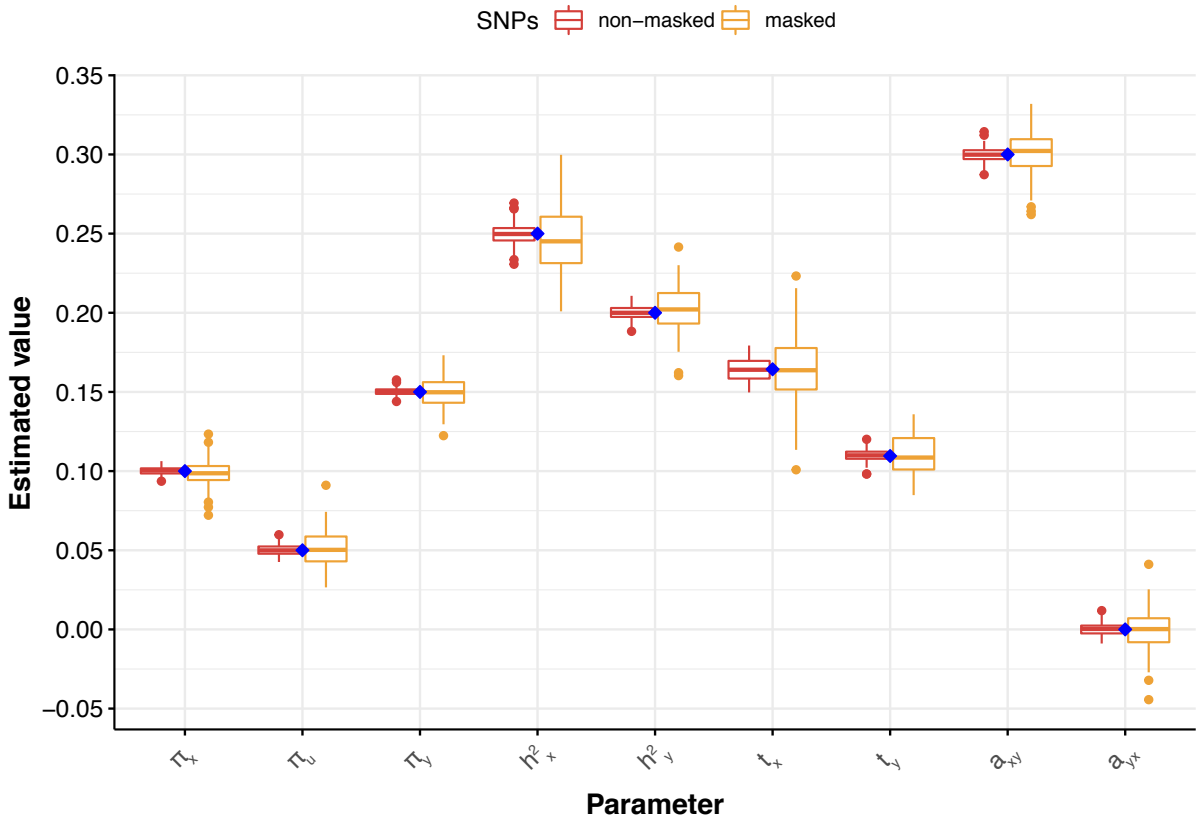


Figure S7: Boxplots representing the distribution of parameter estimates from 100 different data generations. The true values of the parameters used in the data generations are represented by the blue dots. The different coloured boxplots represent the masking of the SNPs during the likelihood optimisation process. Red boxplots are the runs that used the full set of SNPs in the optimisation, whereas orange boxplots had a tenfold masking with only 5,000 SNPs being used in the optimisation.

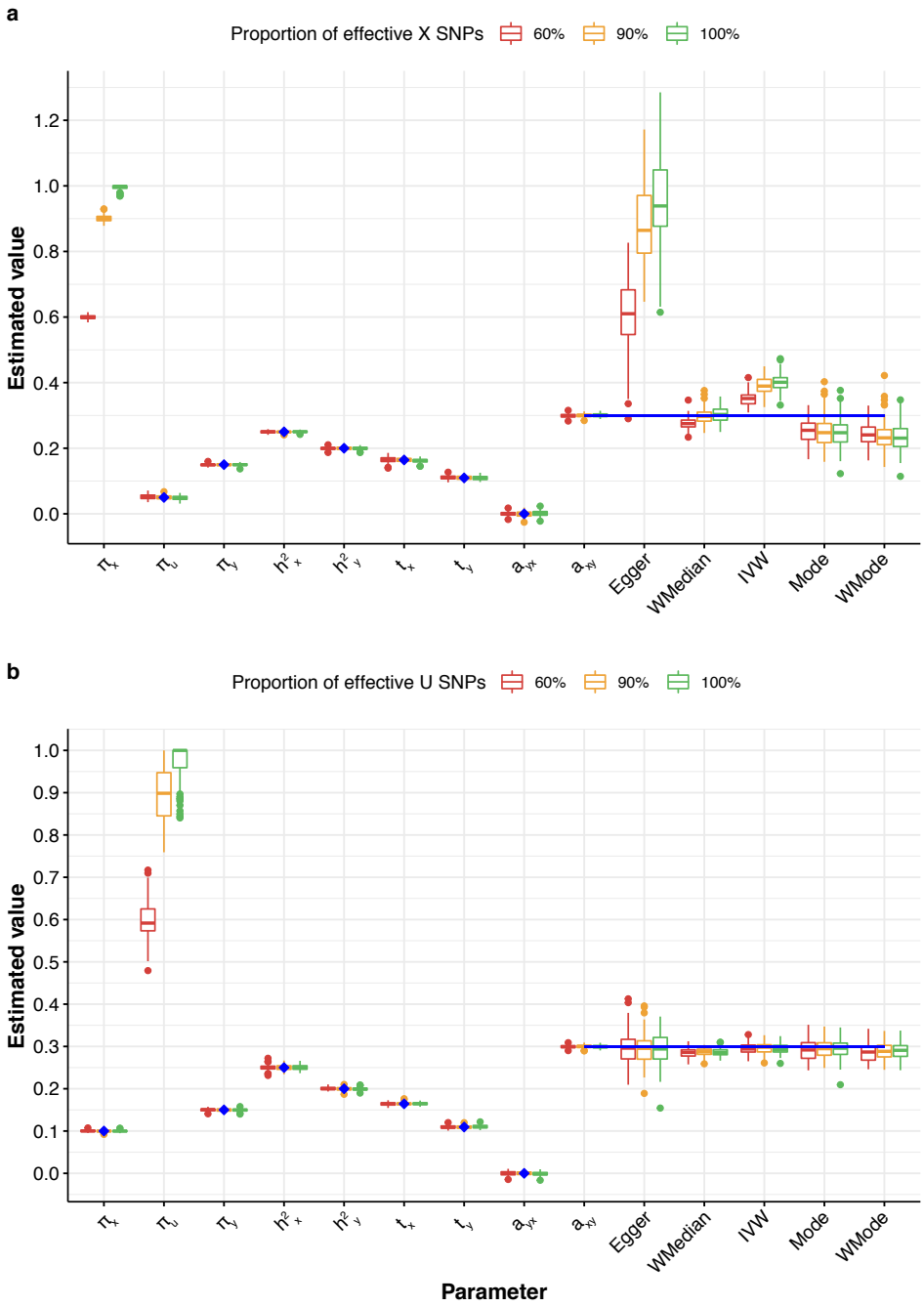


Figure S8: Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. The different coloured boxplots represent the different proportion of effective SNPs that make up the spike-and-slab distributions of the γ vectors. The red boxplots are those when trait X (panel a) or U (panel b) has 60% effective SNPs, the orange ones are when the proportion of effective SNPs for those two is 90%, and the green boxplots is when they have a 100% effective SNPs.

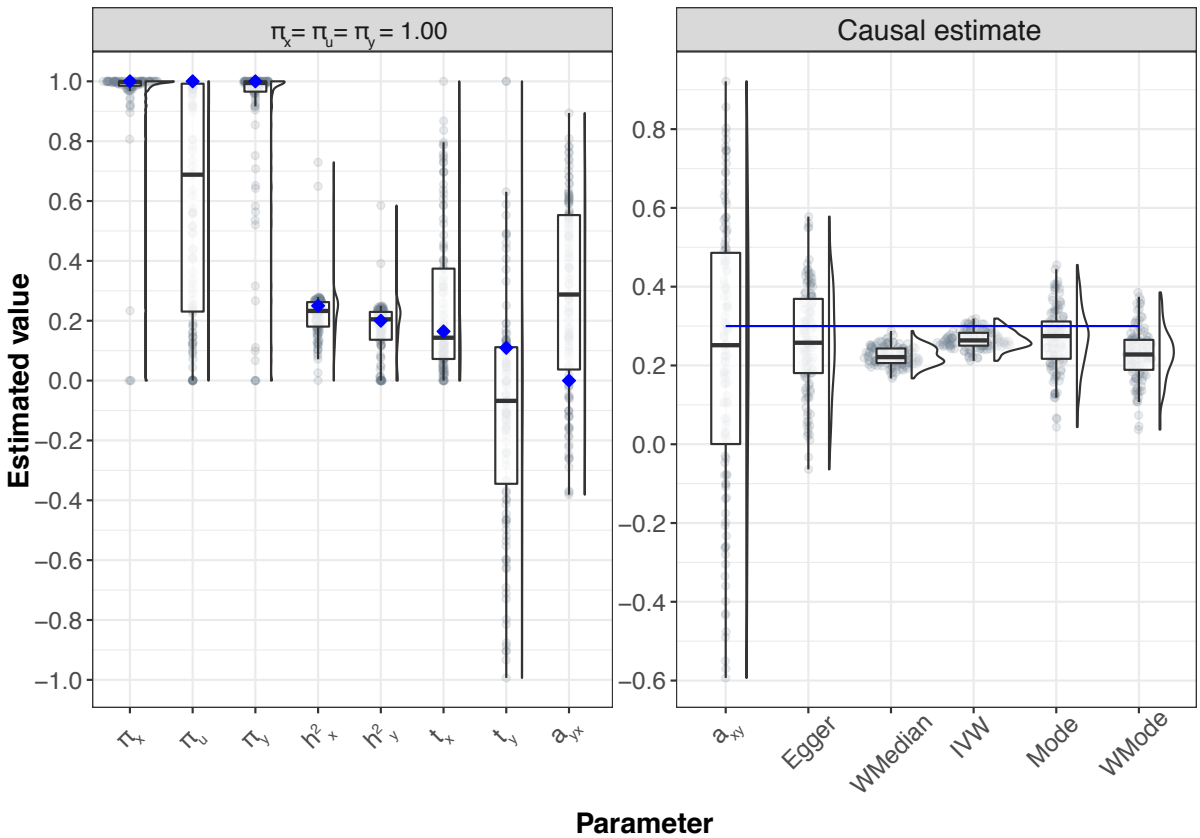


Figure S9: Raincloud boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. The proportion of effective SNPs that make up the spike-and-slab distributions of the γ vectors in this setting is 100% for all traits X, Y and U.

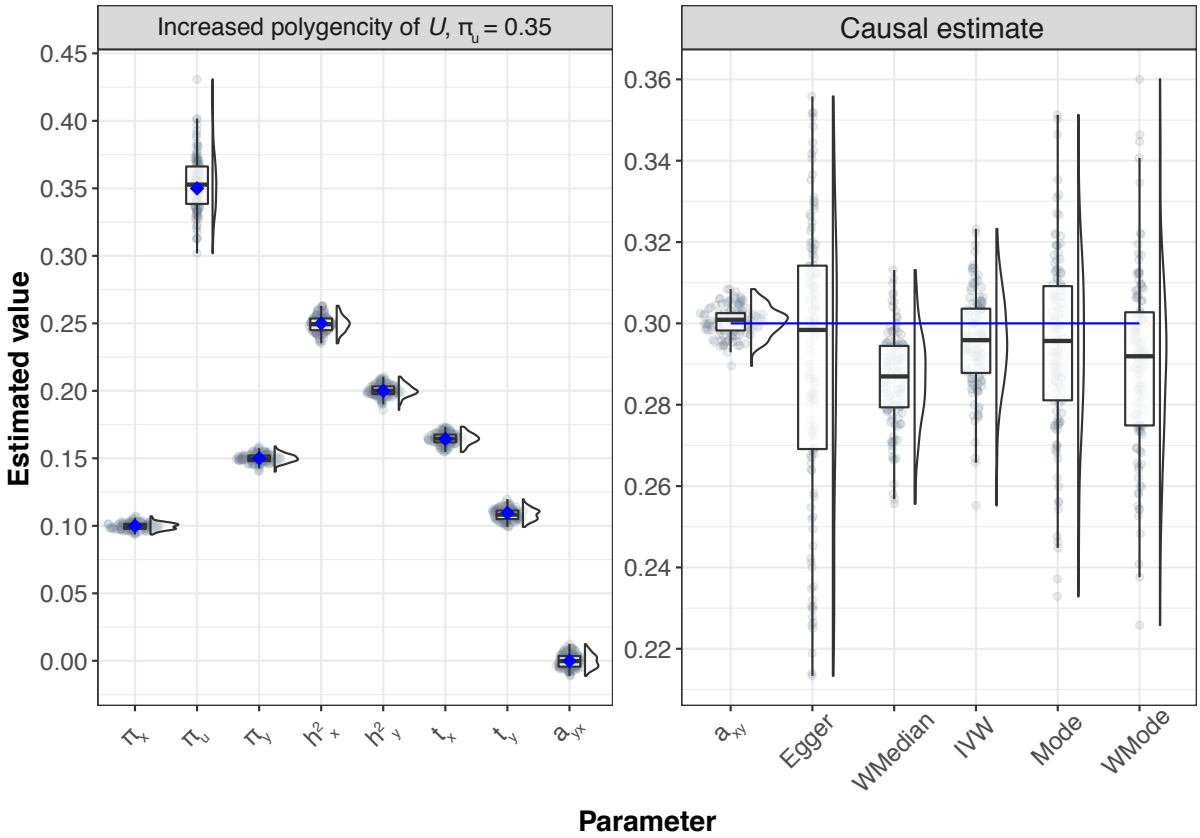


Figure S10: Raincloud boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. In this figure, the polygenicity for U is increased in the form of higher $\pi_u = 0.35$.

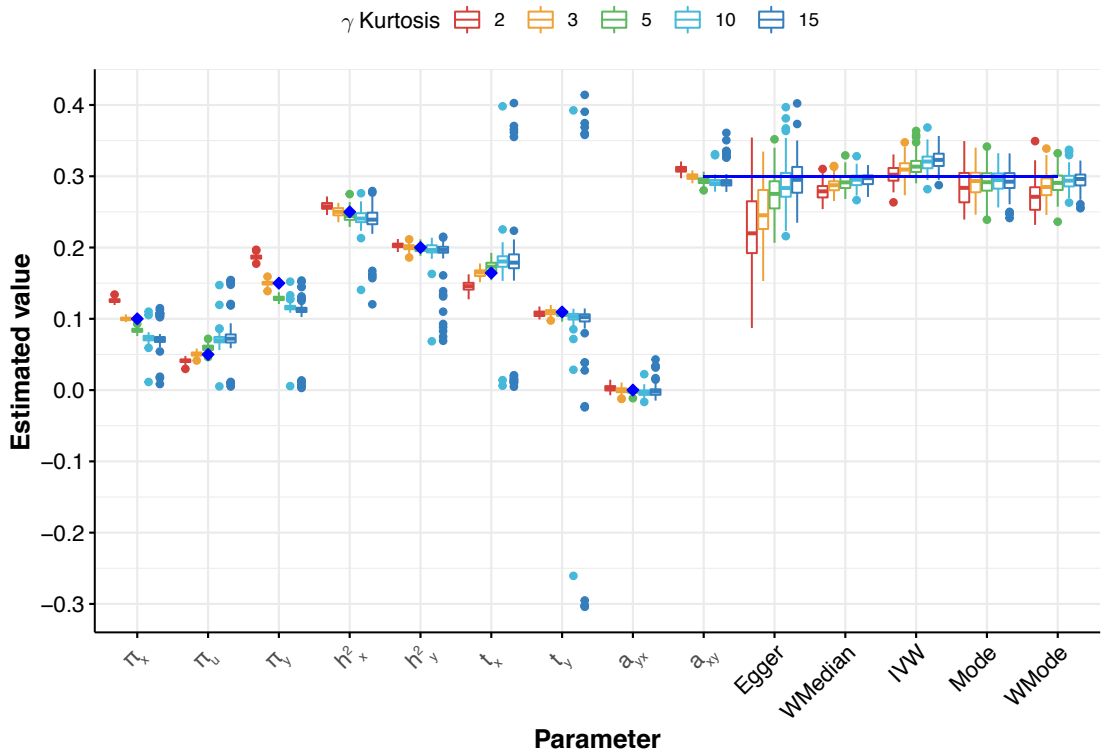


Figure S11: Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue 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.

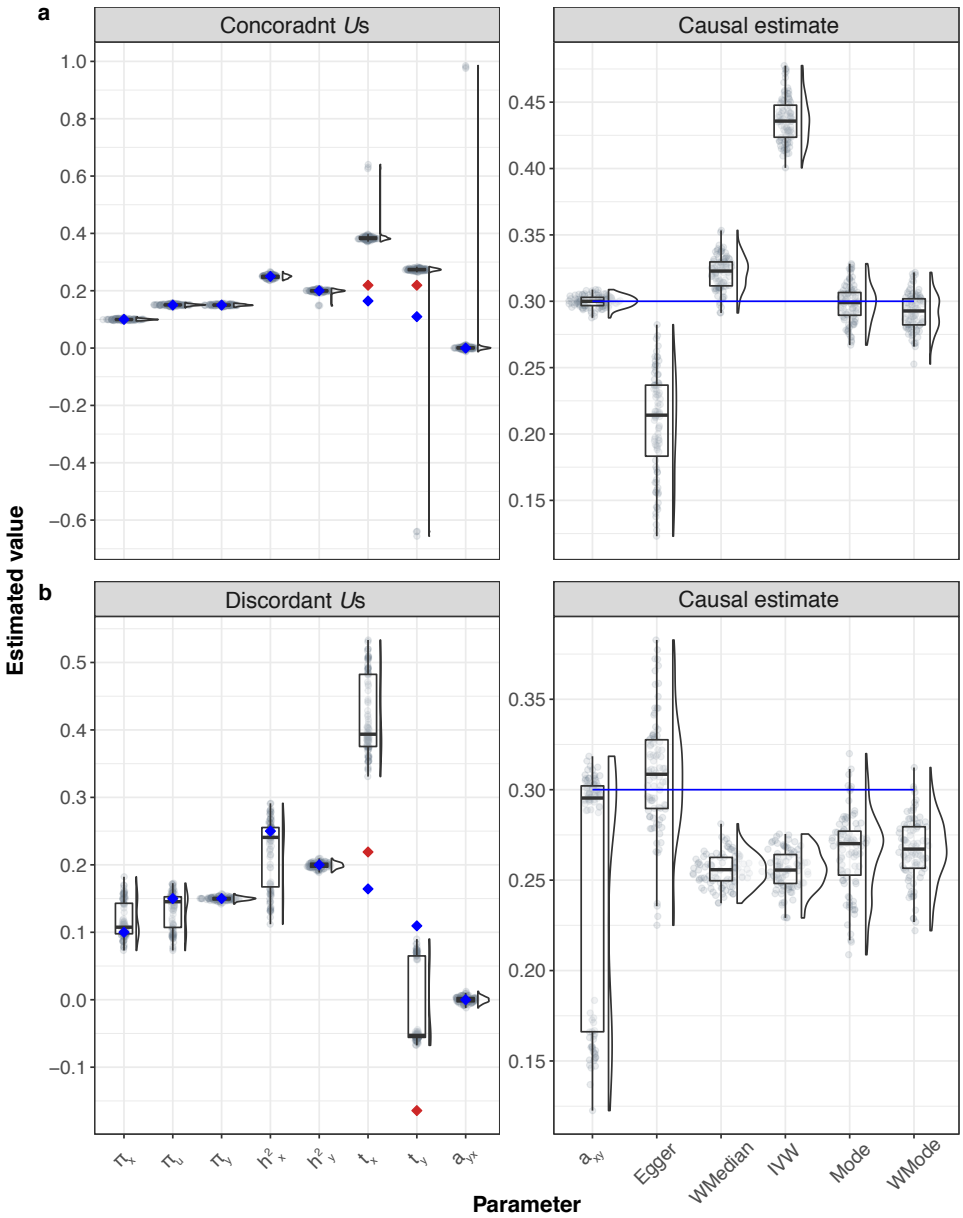


Figure S12: Boxplots representing the distribution of parameter estimates from 100 different data generations. For each generation, standard MR methods as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines. Figure A represents the underlying data generation with two concordant heritable confounders U_1 and U_2 with positive effects on traits X and Y, whereas figure B shows data generations with two discordant heritable confounders with $t_x^{(1)} = 0.16, t_y^{(1)} = 0.11$ shown as blue dots and $t_x^{(2)} = 0.22, t_y^{(2)} = -0.16$ shown as red dots.

Violating the two-component mixture assumption using simulated data

One of the assumptions of LHC-MR is that the direct effects on any of the traits come from a two-component Gaussian mixture. To violate this assumption, we simulated data using a three component mixture with a small number of SNPs having a larger effect on a trait and assessed how LHC-MR performs. We set out to explore whether a third component of effects could force LHC-MR to mask a confounder (if it exists) or would it create an artificial confounder. In addition, we tested whether filtering out SNPs with large effect before fitting the LHC model could remove the third component effect.

Simulations covered two scenarios, (1) introduced a large third component for X while decreasing the strength of U ($\pi_{x1} = 0.1, \pi_{x2} = 0.005, h^2_{x1} = 0.25, h^2_{x2} = 0.2, h^2_u = 0.1$), (2) introduced a large third component for both X and Y while decreasing U ($\pi_{x1} = 0.1, \pi_{x2} = 0.005, h^2_{x1} = 0.25, h^2_{x2} = 0.2, \pi_{y1} = 0.15, \pi_{y2} = 0.002, h^2_{y1} = 0.2, h^2_{y2} = 0.15, h^2_u = 0.1$).For each scenario, the data was simulated 100 times, and the likelihood function optimised to estimate the parameters. Then the process was repeated with SNPs being filtered when their squared effect was greater than an arbitrary-set threshold ($\beta^2 > 0.005/0.001/0.0005$).

Results revealed that the presence of a third component biases parameter estimation only when the introduced third component is very different from the second component (i.e. SNPs acting through the third component having very strong effects), and if the confounder has a moderate effect. It can lead to the selection of a no-confounder model, even when a true U exists (Figure S13 panel a), and slightly bias the causal effect estimates by either merging true effect estimates with confounder effect or leading to a significant estimate of b when no reverse causal effect estimate has been simulated (Figure S13 panel a).

For the first scenario, we observed bimodality for pi_x, pi_u and h^2_x (π_x and h^2_x had one peak corresponding to the values set for the second component), a large overestimation of t_x (third- and possibly part of the second- component acting through U), and t_y estimated as null (non-significant). Reverse causal effect estimate was slightly overestimated and significant (median p-value $= 8.93 \times 10{-}04$).
However, filtering SNPs with strong effects based on a threshold (0.0005 in Figure S13 panel b) before running LHC-MR allowed us to get back correct estimates again. The underestimation of h^2_x is entirely expected after filtering out SNPs strongly associated with X. These results are also reflected in the second scenario (no figure reported).

To assess the existence of a 3rd component in direct effects of real traits, we used mixtures of univariate normal distributions to fit a 3-component model and compare the largest component to the other two. The premise was that if this third component was sufficiently different from the other two, it may represent a third component directly acting on the trait and it can be used to identify a threshold to exclude SNPs that are more likely to belong to this third component (effect size for which the probability to belong to the 3rd component is 10 times larger than for the other two). If the threshold identified is "relatively" small (depends on the genetic architecture of the trait) and if a large proportion of SNPs are excluded based on this threshold, it means that the 3rd component is not too different from the other two, and we could refute the existence of a third component corresponding to direct effects.

We applied this to the thirteen traits that we have, and found that all of them may have a third component for direct effects, with thresholds ranging from 0.011 to 0.033 - but all resulting in the exclusion of less than 0.0014% of the SNPs. We explored how LHC-MR estimates would change if trait pairs that originally had a no-confounder winning model, were filtered for strong-effect SNPs from both the exposure and outcome based on the calculated thresholds. For LDL-BMI, BWeight-BMI and HDL-SVstat, we observed no difference in nesting after filtering

out strong-effect SNPs, in other words, we do not recover any confounder that may have been masked by a strong third component. Thus, we consider the fact that filtering on real traits won't have a big impact as simulation results have shown.

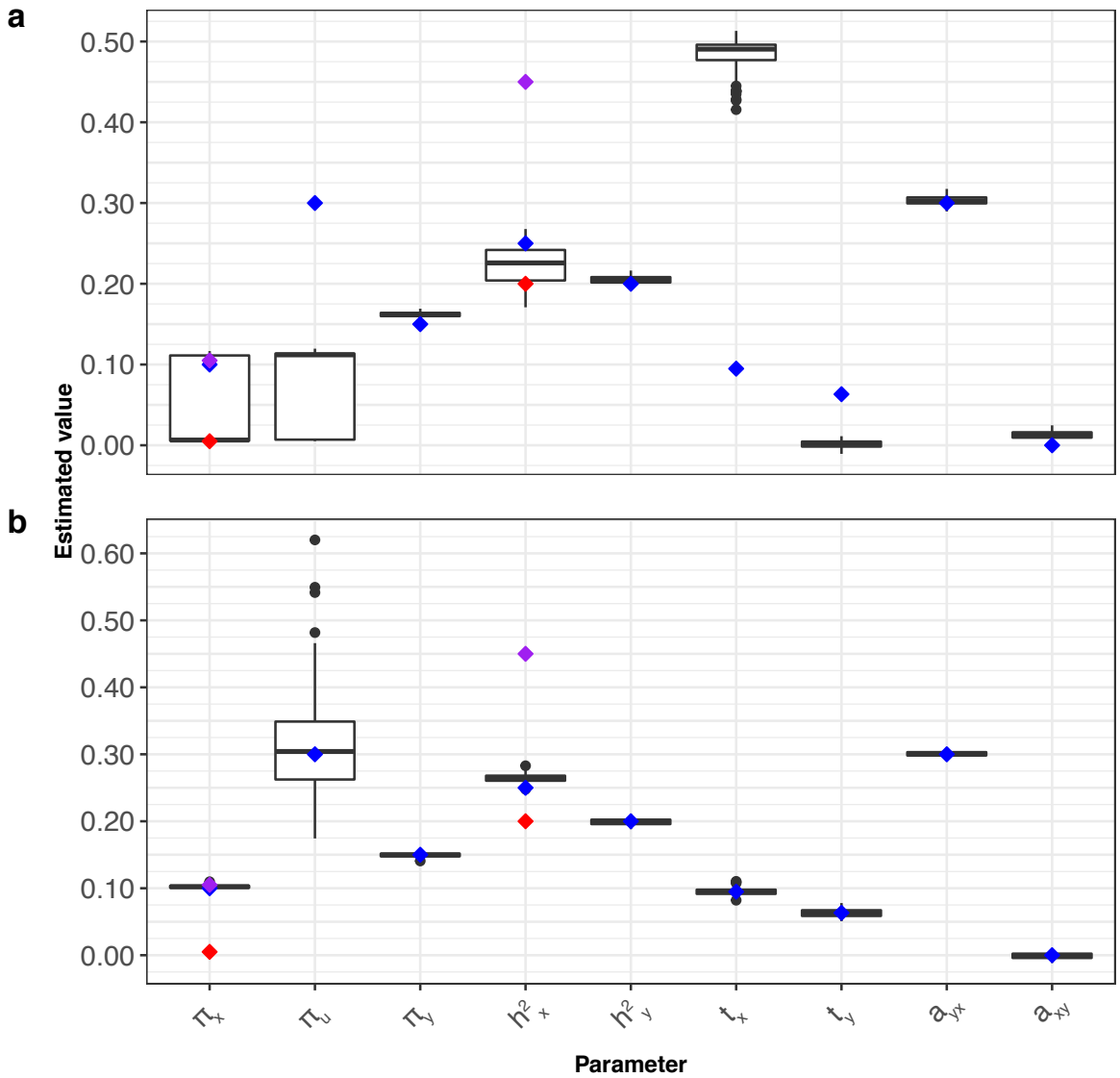


Figure S13: Boxplots of the parameter estimation when a large third component for X is introduced while the strength of U is decreased. Parameter values are in colour, blue for the true values (including second component parameters), red for the third component parameters and purple is used for the parameters that would be obtained by combining the second and third component ($\pi_{x1} = 0.1, \pi_{x2} = 0.005, h^2_{x1} = 0.25, h^2_{x2} = 0.2, h^2_u = 0.1$). Panel a shows the results of estimating parameters using LHC-MR on 100 different simulations, whereas panel b shows the same estimation after filtering out SNPs with a strong effect of X ($b^2 > 0.0005$).

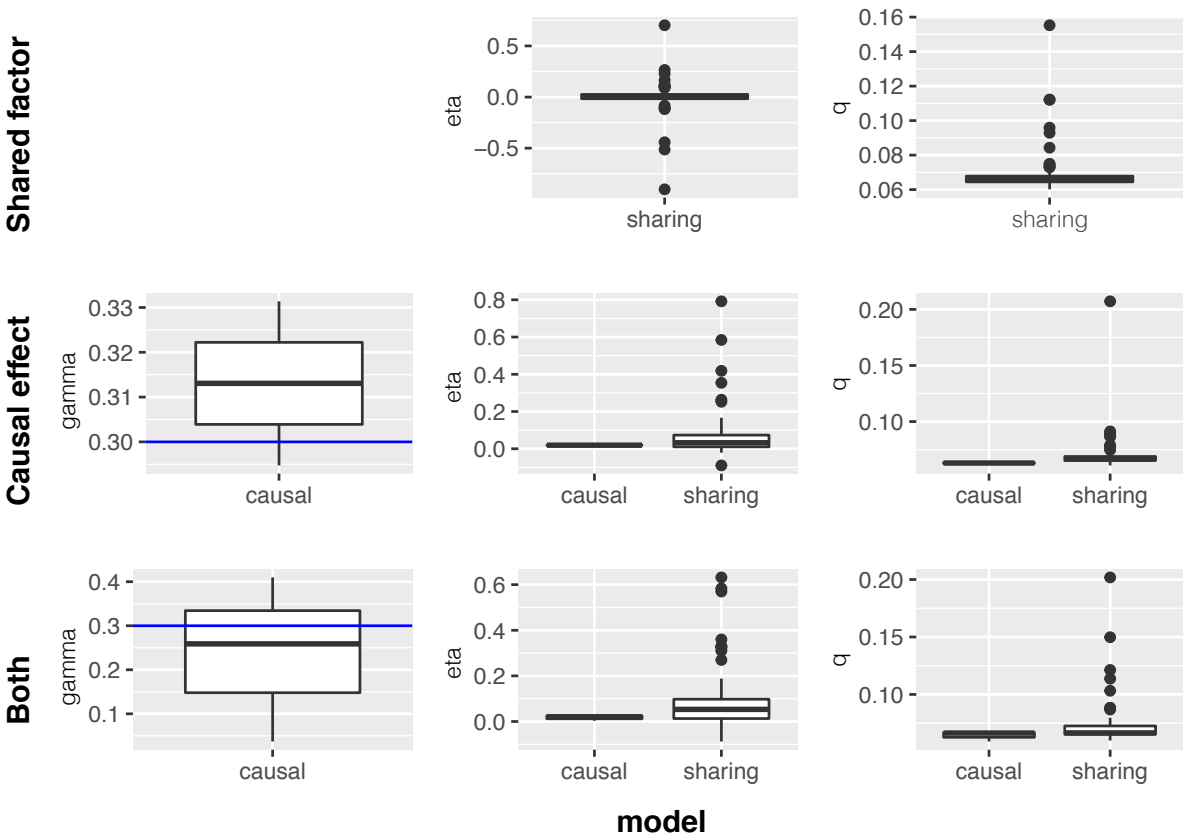


Figure S14: Boxplots of the parameter estimation of CAUSE on LHC-simulated data under three different scenarios: presence of a shared factor only, presence of a causal effect only, presence of both. CAUSE returns two possible models with a respective p-value, the sharing and the causal model, where the causal mode is the significant of the two. When only an underlying shared factor was present in the simulated data, CAUSE had no significant causal estimates. With a true underlying causal effect, the causal model was significant 3% of the simulations. When both an underlying causal effect and a shared factor was present, the causal model was significant only 3% of the simulations.

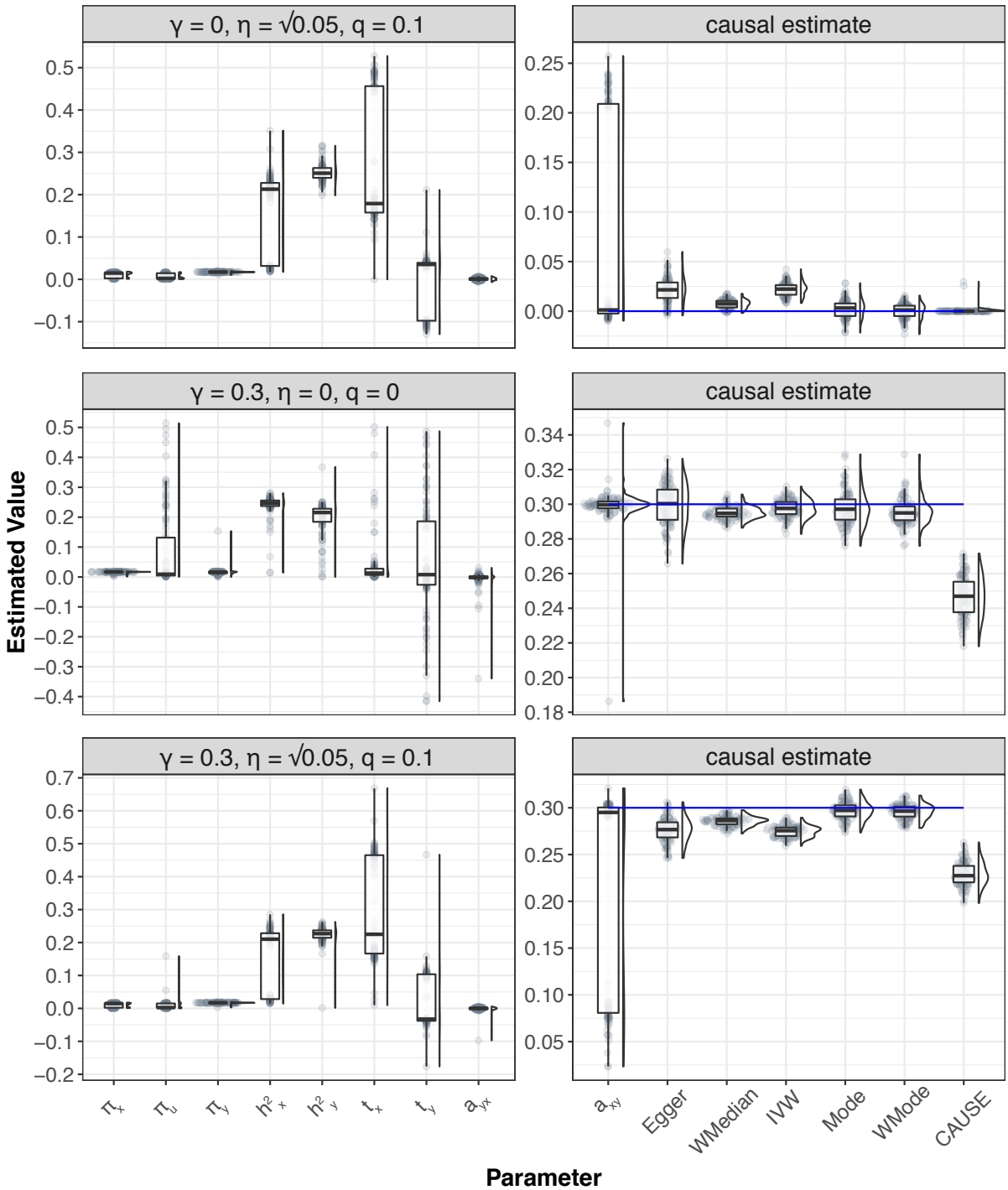


Figure S15: Raincloud boxplots representing the distribution of parameter estimates from LHC-MR of 100 different data generations using the CAUSE framework. For each generation, standard MR methods, CAUSE as well as our LHC-MR were used to estimate a causal effect. The true values of the parameters used in the data generations are represented by the blue dots/lines.

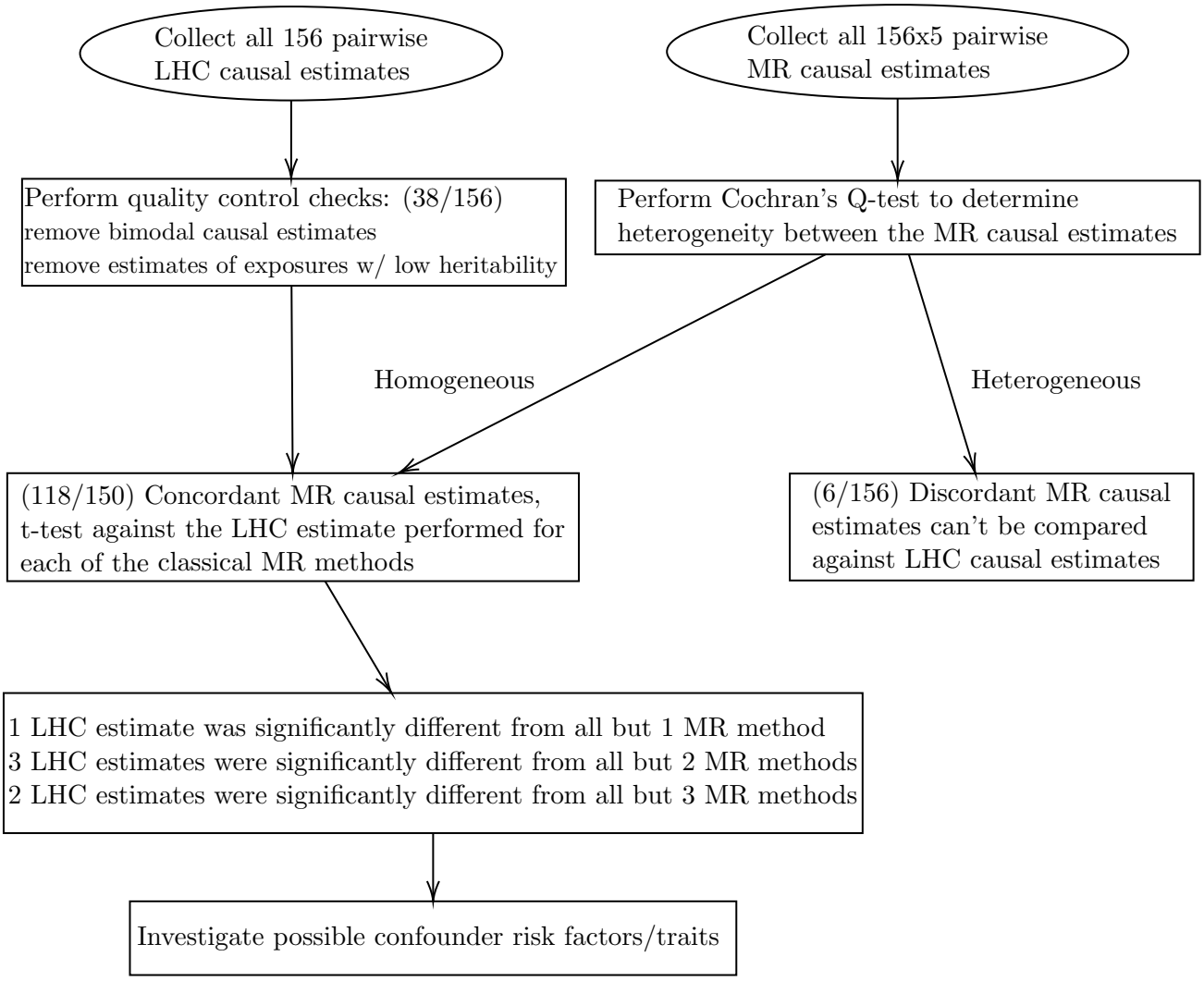


Figure S16: The decision tree used to compare the different causal estimates from standard MR methods to those of LHC.

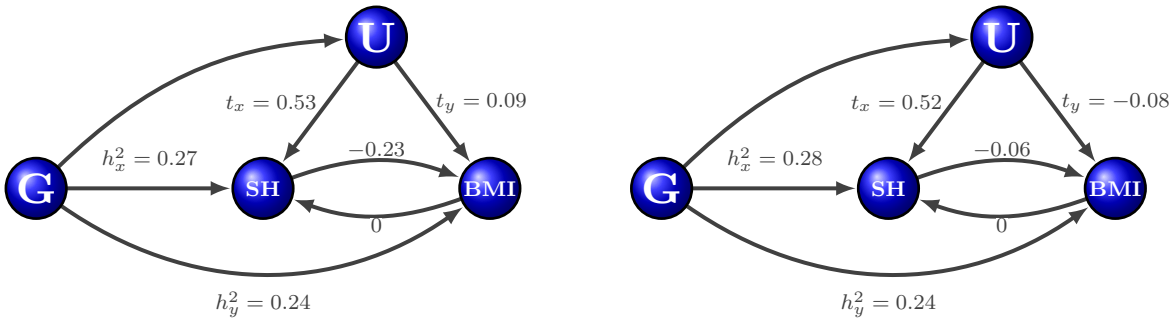


Figure S17: The two equally likely models underlying the observed summary statistics in the case of standing height (SH) and its effect on BMI.

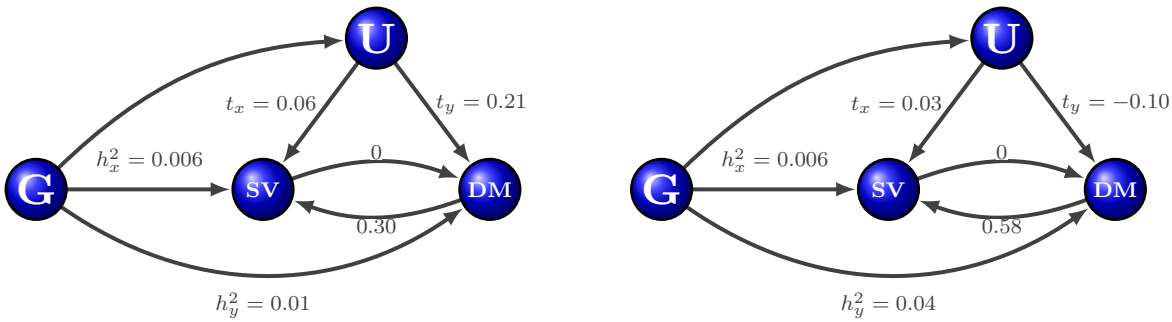
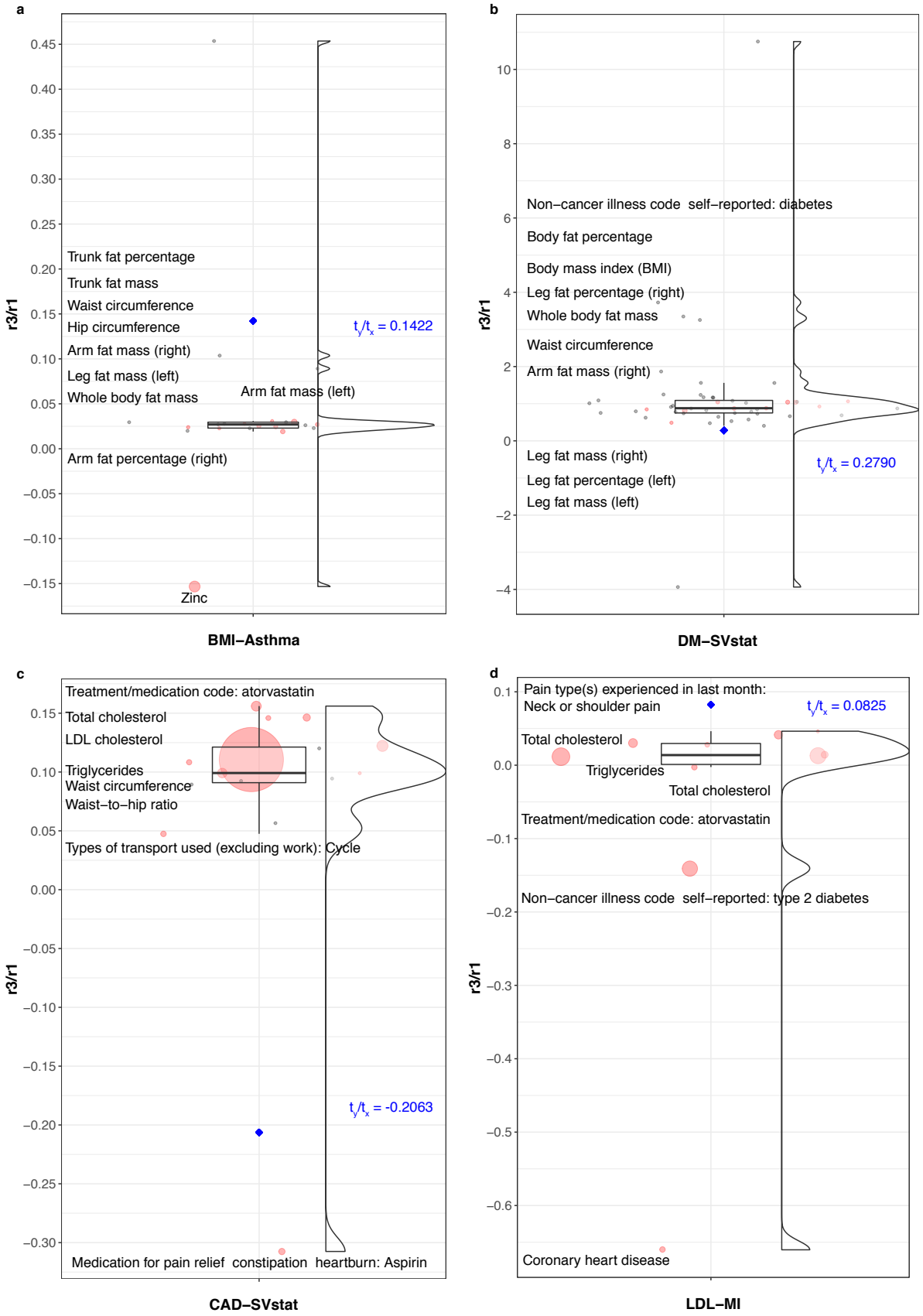
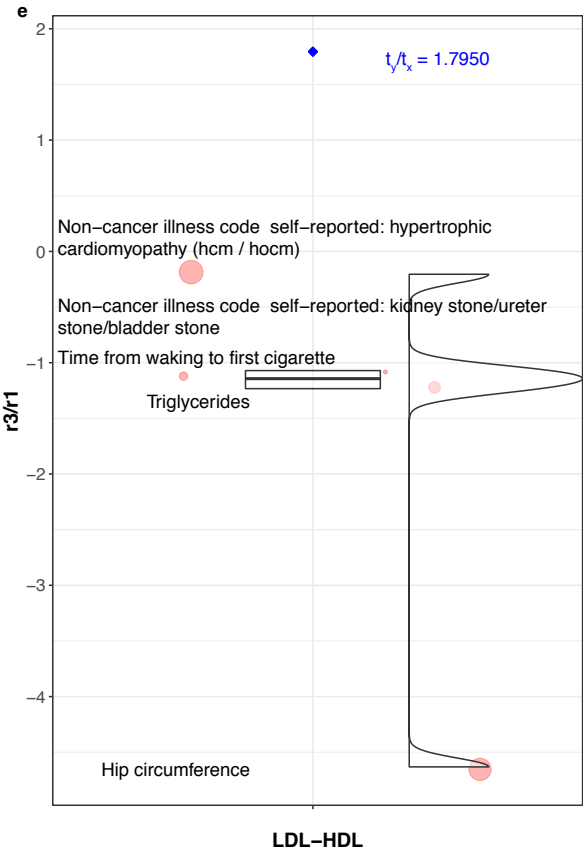
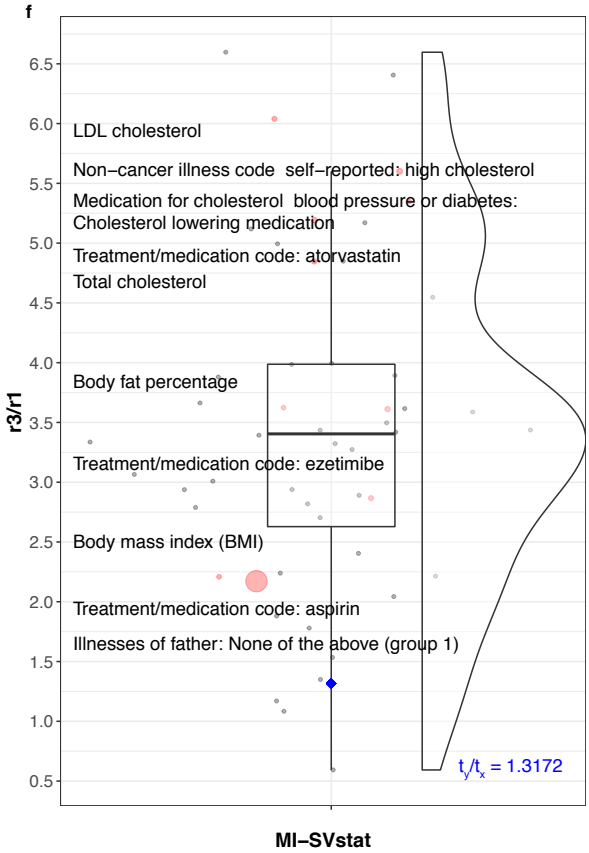
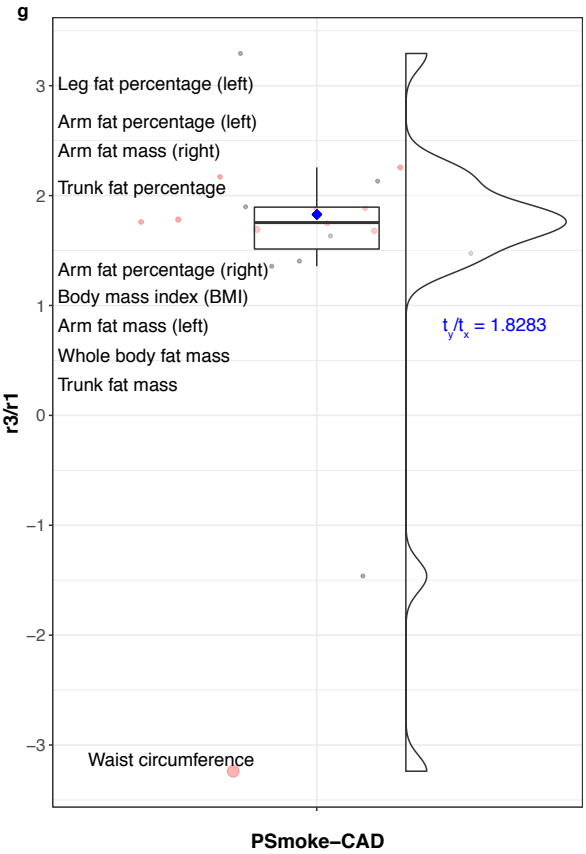


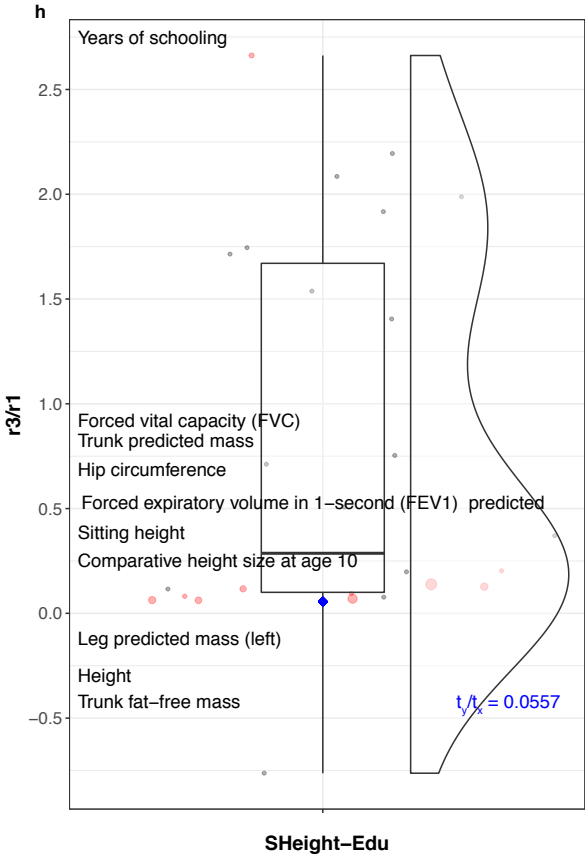
Figure S18: The two equally likely models underlying the observed summary statistics in the case simvastatin (SV/SVstat) on DM.


a
r3/r1
Trunk fat percentage
Trunk fat mass
Waist circumference
Hip circumference
Arm fat mass (right)
Leg fat mass (left)
Whole body fat mass
Arm fat mass (left)
Arm fat percentage (right)
Zinc
t_y/t_x = 0.1422
BMI−Asthma
b
r3/r1
Non−cancer illness code self−reported: diabetes
Body fat percentage
Body mass index (BMI)
Leg fat percentage (right)
Whole body fat mass
Waist circumference
Arm fat mass (right)
Leg fat mass (right)
Leg fat percentage (left)
Leg fat mass (left)
t_y/t_x = 0.2790
DM−SVstat
c
r3/r1
Treatment/medication code: atorvastatin
Total cholesterol
LDL cholesterol
Triglycerides
Waist circumference
Waist−to−hip ratio
Types of transport used (excluding work): Cycle
Medication for pain relief constipation heartburn: Aspirin
t_y/t_x = -0.2063
CAD−SVstat
d
r3/r1
Pain type(s) experienced in last month:
Neck or shoulder pain
Total cholesterol
Triglycerides
Total cholesterol
Treatment/medication code: atorvastatin
Non−cancer illness code self−reported: type 2 diabetes
Coronary heart disease
t_y/t_x = 0.0825
LDL−MI


e
t_y/t_x = 1.7950
Non−cancer illness code self−reported: hypertrophic cardiomyopathy (hcm / hocm)
Non−cancer illness code self−reported: kidney stone/ureter stone/bladder stone
Time from waking to first cigarette
Triglycerides
Hip circumference
r3/r1
LDL−HDL


f
LDL cholesterol
Non−cancer illness code self−reported: high cholesterol
Medication for cholesterol blood pressure or diabetes: Cholesterol lowering medication
Treatment/medication code: atorvastatin
Total cholesterol
Body fat percentage
Treatment/medication code: ezetimibe
Body mass index (BMI)
Treatment/medication code: aspirin
Illnesses of father: None of the above (group 1)
t_y/t_x = 1.3172
r3/r1
MI−SVstat


g
Leg fat percentage (left)
Arm fat percentage (left)
Arm fat mass (right)
Trunk fat percentage
Arm fat percentage (right)
Body mass index (BMI)
Arm fat mass (left)
Whole body fat mass
Trunk fat mass
t_y/t_x = 1.8283
Waist circumference
r3/r1
PSmoke−CAD


h
Years of schooling
Forced vital capacity (FVC)
Trunk predicted mass
Hip circumference
Forced expiratory volume in 1−second (FEV1) predicted
Sitting height
Comparative height size at age 10
Leg predicted mass (left)
Height
Trunk fat−free mass
t_y/t_x = 0.0557
r3/r1
SHeight−Edu

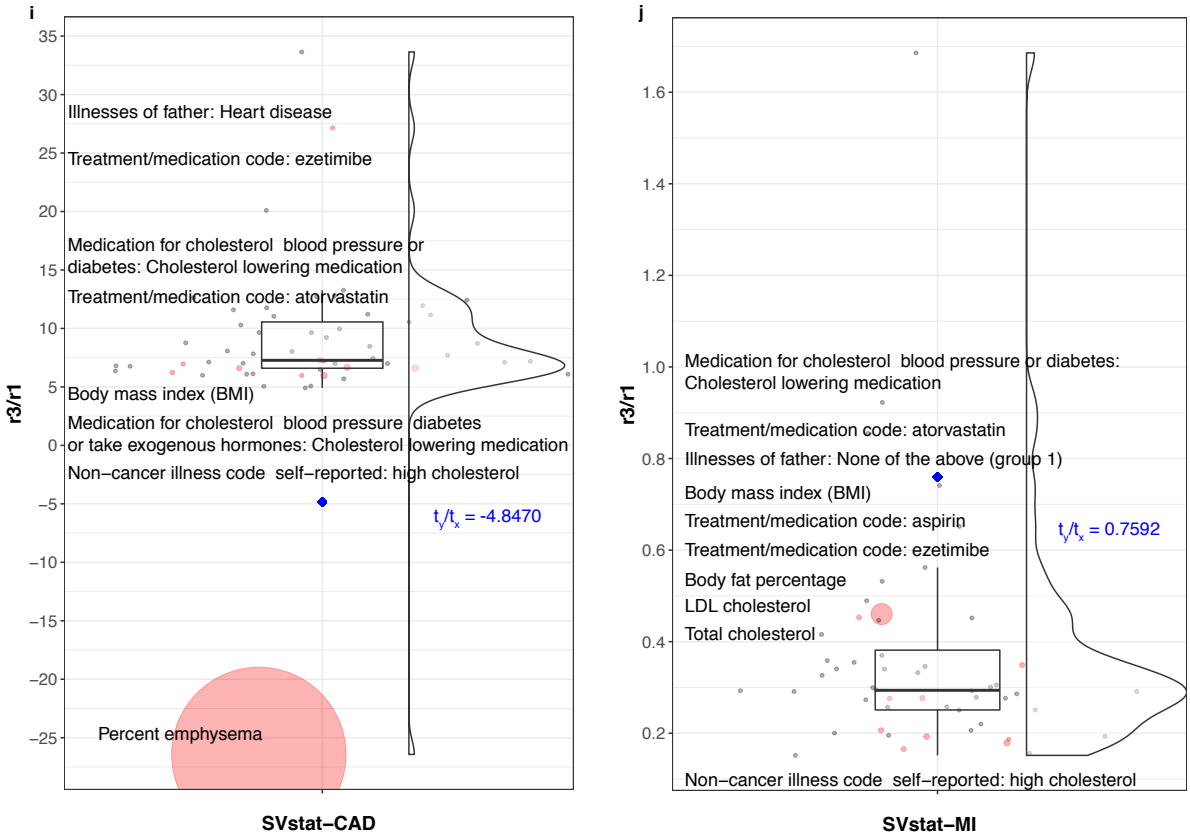


Figure S20: Confounder effects obtained from EpiGraphDB plotted as a raincloud with the r_3/r_1 ratio on the y-axis. The blue diamonds represent the t_y/t_x ratio derived from the LHC model for that trait pair, also reported in blue text. Labelled dots in red and their varying size show the ten largest 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.

a MR-Egger

LHC-MR	Sig+	Sig−	nonSig
Sig+	8	0	10
Sig−	0	1	10
nonSig	1	2	89

b Weighted median

	Sig+	Sig−	nonSig
Sig+	15	0	3
Sig−	0	5	6
nonSig	5	9	78

c IVW

	Sig+	Sig−	nonSig
Sig+	16	0	2
Sig−	0	6	5
nonSig	7	10	75

d Simple mode

LHC-MR	Sig+	Sig−	nonSig
Sig+	10	0	8
Sig−	0	1	10
nonSig	1	0	91

e Weighted mode

	Sig+	Sig−	nonSig
Sig+	10	0	8
Sig−	0	2	9
nonSig	1	4	87

f IVW

Weighted mode	Sig+	Sig−	nonSig
Sig+	11	0	0
Sig−	0	1	5
nonSig	12	15	77

Figure S21: Cross tables between LHC-MR and various standard MR methods comparing the significance and sign of each respective causal estimate. Panel *f* shows a cross table between the two-least correlated MR methods in terms of their estimates.

Supplementary Tables

Sample Size	LHC	MR Egger	IVW	Mode	Weighted median	Weighted mode
10,000	0.4	4.44	0.27	0.29	0.27	0.28
50,000	0.11	0.26	0.08	0.1	0.1	0.11
100,000	0.07	0.12	0.05	0.07	0.07	0.08
200,000	0.04	0.07	0.02	0.04	0.04	0.05
300,000	0.01	0.06	0.01	0.03	0.03	0.04
400,000	0.01	0.06	0.01	0.03	0.02	0.03
500,000	0.00	0.06	0.02	0.02	0.02	0.02

Table S1: Table reporting the RMSE in estimating the causal effect using different MR methods and LHC-MR with varrying sample sizes.

UKBB ID / Data Origin	Trait Name	Abbreviation	Sample Size	PMID
845	Age completed full time education	Edu	240,547	25826379
21001_irnt	Body mass index (BMI)	BMI	359,983	25826379
2443	Diabetes diagnosed by doctor	DM	360,192	25826379
20002_1075	Non-cancer illness code, self-reported: heart attack/myocardial infarction	MI	361,141	25826379
20002_1111	Non-cancer illness code, self-reported: asthma	Asthma	361,141	25826379
2887	Number of cigarettes previously smoked daily	PSmoke	84,456	25826379
20022_irnt	Birth weight	BWeight	205,475	25826379
50_irnt	Standing height	SHeight	360,388	25826379
4080	Systolic blood pressure, automated reading	SBP	340,159	25826379
20003_1140861958	Treatment/medication code: simvastatin	SVstat	361,141	25826379
GLGC	LDL Cholesterol	LDL	89,138	24097068
GLGC	HDL Cholesterol	HDL	93,561	24097068
UKBB + CARDIoGRAMplusC4D	Coronary Artery Disease	CAD	380,831	29212778

Table S2: Details of the origin study of each trait, its abbreviation used in this paper, the sample size of the study for that trait, as well as the PubMed article ID.

Exposure	Outcome	Ratio lower_CI	Ratio mean	Ratio upper_CI
Edu	BMI	0.8472	0.9160	1.0713
SHeight	Edu	-0.9816	0.8538	3.0512
SHeight	BMI	0.6388	0.8252	1.1771
CAD	SVstat	0.6576	0.6941	0.7287
SBP	SHeight	0.6162	0.6880	0.7609
CAD	HDL	0.2505	0.3749	0.5266
SVstat	DM	0.1558	0.1905	0.2290
CAD	DM	-0.0499	0.1873	0.2866
SBP	DM	0.0189	0.0688	0.1078

Table S3: Table reporting the proportion of the genetic covariance attributed to the bi-directional causal effect. For each trait pair, parameter estimates from LHC-MR were used to calculate those two values from Eq. 2, by sampling 1,000 times from distributions provided by LHC-MR (parameter estimates and covariance matrices). The mean of the 1,000 computed ratios, as well as the lower and upper bounds of the 95% confidence interval (CI) are reported above.