Welch-weighted Egger regression reduces false positives due to correlated pleiotropy in Mendelian randomization

Brielin C. Brown*[1, 2] and David A. Knowles†[1, 2, 3, 4]

[1]*Data Science Institute, Columbia University, New York, NY 10027*
[2]*New York Genome Center, New York, NY 10013*
[3]*Department of Computer Science, Columbia University, New York, NY 10027*
[4]*Department of Systems Biology, Columbia University, New York, NY 10027*

Abstract

Modern population-scale biobanks contain simultaneous measurements of many phenotypes, providing unprecedented opportunity to study the relationship between biomarkers and disease. However, inferring causal effects from observational data is notoriously challenging. Mendelian randomization (MR) has recently received increased attention as a class of methods for estimating causal effects using genetic associations. However, standard methods result in pervasive false positives when two traits share a heritable, unobserved common cause. This is the problem of correlated pleiotropy. Here, we introduce a flexible framework for simulating traits with a common genetic confounder that generalizes recently proposed models, as well as simple approach we call Welch-weighted Egger regression (WWER) for estimating causal effects. We show in comprehensive simulations that our method substantially reduces false positives due to correlated pleiotropy while being fast enough to apply to hundreds of phenotypes. We first apply our method to a subset of the UK Biobank consisting of blood traits and inflammatory disease, and then a broader set of 411 heritable phenotypes. We detect many effects with strong literature support, as well as numerous behavioral effects that appear to stem from physician advice given to people at high risk for disease. We conclude that WWER is a powerful tool for exploratory data analysis in ever-growing databases of genotypes and phenotypes.

1 Introduction

Modern population-scale biobanks contain genetic information with simultaneous measurements of many phenotypes, providing unprecedented opportunity to study the relationship between biomarkers and disease. However, inferring causal effects from observational data is notoriously challenging. Mendelian randomization (MR) has recently received increased attention as a class of methods that can mitigate issues in causal inference by using genetic variants (single nucleotide polymorphisms, SNPs) from genome-wide association studies (GWAS) as instrumental variables to determine the causal effect of an exposure (A) on an outcome (B). To estimate causal effects, MR methods must make strong assumptions that limit their ability to be applied at scale. Perhaps the most problematic assumption is that the SNP only affects B through A, *i.e.* there is no horizontal pleiotropy. Recent methods such as Egger regression and the mode-based-estimator are able to relax this assumption, instead assuming there is no correlated horizontal pleiotropy or modal pleiotropy, respectively [1, 2]. Correlated horizontal pleiotropy arises when both A and B share a common heritable factor (U in Figure 1A), resulting in genetic correlation between the traits in the absence of a causal

*Corresponding author, bb2991@columbia.edu

†Co-corresponding author, dak2173@columbia.edu

effect. This kind of pleiotropy is both challenging to handle and thought to be pervasive between traits that share underlying biological processes.

Recently, computationally-intensive mixture models such as CAUSE [3] and MRMix [4] have shown success at estimating causal effects in the presence of correlated pleiotropy. These approaches are similar in that they both assume that a proportion of the instruments are valid, acting only on the exposure, and a proportion are invalid with correlated effects on the exposure and outcome that are not mediated by a causal effect. However, these approaches differ in the way they model the data-generating distribution. CAUSE explicitly models a latent factor, assuming that direct and latent factor-mediated genetic effects on the exposure have the same per-variance effect distribution and that the latent factor has a smaller effect on the outcome than the exposure. The MRMix model instead assumes the effects of invalid instruments come from a bivariate normal without explicitly modeling the latent factor. Their estimation methods also differ, with CAUSE using Bayesian model comparison and MRMix attempting to maximize the number of instrument effects lying close to the inferred causal effect. Another approach, the latent causal variable (LCV) model, is able to detect causality under arbitrarily-structured pleiotropy [5]. This model assumes all genetic correlation is mediated by the effect of the latent variable, with "causality" when the latent variable is highly genetically correlated with the exposure. The LCV method then estimates the "genetic causality proportion" (GCP), with larger values indicating the exposure is more likely to be "genetically casual" for the outcome. However, GCP is not directly interpretable as a causal effect size.

Most MR studies presuppose the direction of effect, specifying one phenotype as the outcome and the other as the exposure. Pre-specifying the effect direction can be sound when the outcome is clearly biologically downstream of the exposure, but many cases are less clear-cut and it is preferable to learn the direction of the effect from the data. Some researchers have instead explored bidirectional MR [6, 7], which tests for an effect in each direction, or gwas-pw [8], which infers the effect direction from the data. Others have used Steiger filtering to remove instruments that might be acting on the outcome, rather than the exposure, which has been shown to reduce false positives due to misspecification of the exposure-outcome relationship [9]. However, the utility of these approaches for complex traits, which might contain non-causal correlated pleiotropy, is questionable [5].

Here, we introduce a flexible model for bi-directional MR that explicitly models the genetic architecture of both the observed phenotypes and a heritable confounder while allowing for arbitrary linear dependencies between them (Figure 1A). Our model captures recently proposed models for the MR data generating process, including those used in LCV [5] and CAUSE [3], as special cases. We also introduce a simple method for producing causal effect estimates that leverages a secondary dataset to filter and down-weight likely pleiotropic SNPs in an Egger-like regression, an approach we call Welch-weighted Egger regression (WWER, Figure 1C-D). By filtering SNPs with indistinguishable statistical effects on the exposure and the outcome, our method can be seen as an extension of Stieger filtering. To our knowledge Steiger filtering has not been extensively evaluated as an approach to dealing with non-causal association due to correlated pleiotropy. We show via extensive simulations varying the trait and model architectures that our approach reduces false positives due to correlated pleiotropy while being computationally efficient enough to apply in bi-directed exploratory data analyses of hundreds of phenotypes. We first apply our method to a limited set of phenotypes from the UK Biobank (UKBB) consisting of blood biomarker and blood cell composition traits, as well as common inflammatory diseases, and recover signals corresponding to known disease risk factors. We next apply our method broadly to over 400 phenotypes from the UKBB, again recovering known disease risk factors, while also finding broad signatures of risk factors on behavior, likely reflecting patient response to common medical advice.

2 Methods

Bi-directed Mendelian randomization model

We introduce a flexible model that allows for both unidirectional and bidirectional causal effects while explicitly modeling the genetic architecture of each trait. In contrast to previously proposed models [5, 3, 4],

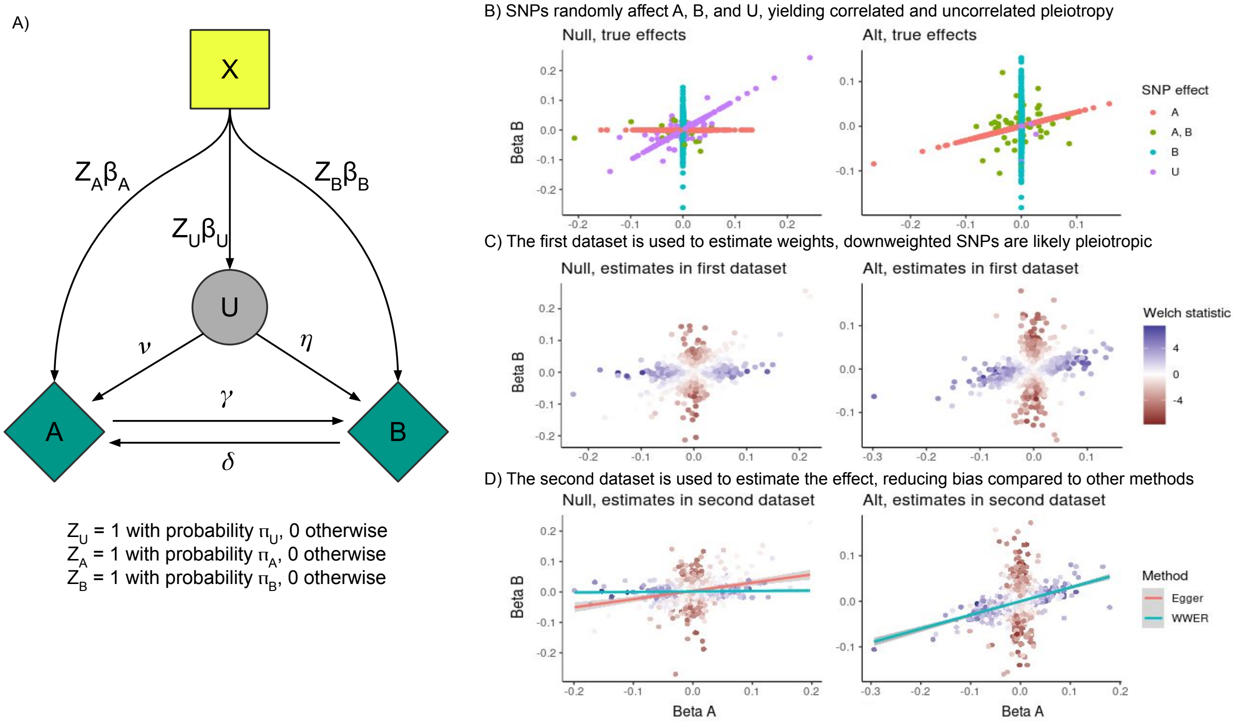


Figure 1: Our model for bi-directional Mendelian randomization, along with an example demonstrating the utility of WWER as compared to standard Egger regression under both the null and 1-way alternative hypothesis. A) A flexible model for bi-directional Mendelian randomization. SNPs X can affect the unobserved phenotype (confounder) U as well as the observed phenotypes of interest A and B. η and ν represent the per-variance effect of U on A and B, respectively, while γ and δ represent the per-variance causal effect of A on B and B on A, respectively. The allelic architecture of each phenotype can be independently adjusted by adjusting the proportion of effect variants, π, and variance of the distribution of effect sizes, β. B) The true effect of each SNP on phenotype A vs B under (left) a null model with $\eta = \nu = \sqrt{0.2}$ and $\gamma = \delta = 0$ and (right) an alternative (alt) model with $\eta = \nu = \delta = 0$ and $\gamma = \sqrt{0.2}$. Our method uses two independent samples to estimate causal effects. C) In the first sample, WWER calculates the Welch statistic, with large positive values (blue) indicating the SNP has a stronger effect on A and large negative values (red) indicating the SNP has a stronger effect on B. SNPs with near-equal effects on the diagonal axis get scores near 0. D) In the second sample, WWER filters SNPs with low Welch statistic, then uses the Welch statistic as a weight for the remaining SNPs when regressing the effect of the outcome on the exposure. Under the null (left) Egger regression produces a false positive, whereas WWER down-weights pleiotropic SNPs and does not. Under the alternative (right) both methods produce nearly-identical results.

we explicitly model the genetic architecture of the confounder and decouple it from that of the exposure, allowing for arbitrary linear effects of the confounder on the pair of observed phenotypes. Our model is also agnostic to the labeling of either observed phenotype as the exposure or the outcome. For this reason, we use A and B to denote the observed traits in the study, and U to denote the unobserved genetic confounder. SNPs X affect each of A, B and U with probabilities q, r, s and effect sizes $\beta_A, \beta_B, \beta_U$ sampled from a normal distribution with variances $\sigma_A^2, \sigma_B^2, \sigma_U^2$, respectively. The probability of effect and variance of the sampling distribution combine to determine the genetic architecture of each trait independently of the others. Finally, η and ν specify the effect of the hidden confounder U on A and B, while γ and δ model the causal effect of A on B and B on A, respectively. Under this model, the phenotype values are given by

$$U = \boldsymbol{X}\beta_U \circ Z_U + \epsilon_U, \tag{1}$$

$$A = U\nu + B\delta + \boldsymbol{X}\beta_A \circ Z_A + \epsilon_A, \tag{2}$$

$$B = U\eta + A\gamma + \boldsymbol{X}\beta_B \circ Z_B + \epsilon_B, \tag{3}$$

where Z's represent indicator variables that the SNP affects that trait, sampled as indicated above, $\circ$ indicates vector element-wise (Hadamard) multiplication, and bolding represents matrices. The recursive nature of this model makes sampling from it non-trivial. In Supplemental Methods we describe how to simulate from this model and how to parameterize the model in terms of the heritability of each phenotype rather than the variance of the effect size distribution. We also explicitly describe how to set the parameters to mimic the models considered in [5] and [3].

Estimation procedure

To produce effect estimates, we introduce a simple method based on a modification to Egger regression [1] that down-weights likely pleiotropic SNPs. Similarly to Steiger filtering [9], we leverage the intuition that if A causes B and a SNP affects A directly, the per-variance effect of the SNP on B can be no larger than the per-variance effect of the SNP on A times the per-variance effect of A on B. That is, the SNP must have its per-variance contribution to B reduced by the effect of A on B. We use this to construct an alternative weighting scheme for Egger regression. First, we select a p-value threshold p_t (usually 5×10^{-8}). For both phenotypes A and B, we construct a set of marginally associated SNPs at threshold p_t. For this set of SNPs, we calculate a weight based on the Welch test statistic for a two-sample difference in mean with unequal variances, and the standard inverse-variance weight. If $\hat{\beta}_{A,k}$ and $\hat{\beta}_{B,k}$ are our estimates of the effect of SNP k on phenotypes A and B, respectively, with $\hat{s}_{A,k}$ and $\hat{s}_{B,k}$ their standard errors, the Welch test statistic [10] is

$$t_k = \frac{|\hat{\beta}_{A,k}| - |\hat{\beta}_{B,k}|}{\sqrt{\hat{s}_{A,k}^2 + \hat{s}_{B,k}^2}} \tag{4}$$

and our weight is

$$w_k = \begin{cases} \frac{t_k}{\bar{t}\hat{s}_{A,k}^2} & \text{if } |t| > t_{min}, \\ 0 & \text{if } |t| \leq t_{min} \end{cases} \tag{5}$$

where $\bar{t}$ is the mean Welch statistic, and t_{min} is the SNP inclusion threshold. We use these SNP weights in the Egger regression of B on A and vice versa for the reverse direction. To avoid bias, we must use two sets of summary statistics. The first set is for SNP selection and weight construction, and the second set is for estimating the causal effect. This method has two parameters, p_t and t_{min}. Here we choose not to tune them and instead always set them to $p_t = 5 \times 10^{-8}$, corresponding to genome-wide significance, and $t_{min} = 1.96$, corresponding to a two-sided p-value for a difference in mean effect of 0.05.

Simulation strategy

We assessed the calibration of WWER under the two-way null as compared to other methods under a broad range of simulation settings. In total we simulated 82 different combinations of simulation parameters in our model. The first 20 settings are designed to mimic the simulations in the LCV study [5], the next 20 settings are designed to mimic the simulations in the CAUSE study [3], and the final 42 explore various combinations of high and low polygenicity for each of the observed and unobserved traits. In all cases we evaluate the false positive rate (FPR) in both the A to B direction and the B to A direction. We also calculate the mean absolute error (MAE) of the effect size estimate. We compare WWER to the standard methods of inverse variance weighting (IVW) and Egger regression, as well as several more recently proposed methods: CAUSE [3], MRMix [4], MR PRESSO [11], raps [12], the weighted median estimator (WME) [13] and the mode-based estimator (MBE) [2]. These methods were chosen as they represent recent substantial contributions to the MR literature with varying estimation procedures and assumptions. For an overview of the approaches and assumptions of each of these methods see Table 1. We also compare against Egger regression with Steiger filtering [9], which has not previously been evaluated for the purpose of handling correlated pleiotropy in bi-directional MR. We intentionally excluded methods such as gwas-pw [8] and LCV [5] that cannot produce bi-directed effect estimates.

Method	Description	Key Assumptions
Egger regression [1]	Regresses genetic effects on outcome on genetic effects on expsure, with an intercept term accounting for uncorrelated pleiotropy.	No correlated pleiotropy (Instrument Strength Independent of Direct Effect, InSIDE).
IVW	An inverse-variance weighted meta-analysis (across variants) of ratio estimates.	InSIDE. Pleiotropic effects have mean 0 (balanced pleiotropy).
MR PRESSO [11]	An outlier removal approach with several diagnostic tools.	Invalid instruments are outliers, InSIDE.
raps [12]	A profile likelihood approach designed to reduce weak instrument bias while handling outliers.	Invalid instruments are outliers, InSIDE.
Mode-based [2]	Takes the mode of the smoothed empirical density of ratio estimates as the causal effect estimate.	ZEro Modal Pleiotropy Assumption (ZEMPA): the most common causal effect estimate is a consistent estimate of the true causal effect.
WME [13]	Causal estimate is weighted median of ratio estimates.	Majority of instruments are valid.
LCV [5]	Assumes all correlation is mediated by shared latent factor. Calculates a "genetic causality proportion" where high values indicate the latent factor is highly genetically correlated to the exposure.	Joint effect size distribution for traits is a sum of 1) shared genetic component due to heritable latent factor and 2) arbitrary distribution not contributing to the genetic correlation.
MRMix [4]	Fits a normal mixture model of genetic effect sizes where a fraction of variants affecting the exposure are valid, and the remainder affect either the outcome, both traits, or neither trait.	ZEMPA. Balanced pleiotropy. Genetic effect size follow normal mixture distribution.
CAUSE [3]	Fits a Bayesian Gaussian mixture model where SNPs affect either i) the exposure, ii) a shared factor or iii) neither	$<50\%$ of variants act through the confounder. Confounder has a stronger effect on exposure than outcome.
Steiger [14]	Filters instruments that are likely to be acting on the outcome rather than the exposure.	No horizontal pleiotropy.
WWER	Uses a training dataset to construct SNP weights that reduce the influence of pleiotropic effects.	Equal trait genetic architecture between training and test studies. Existence of detectable genetic instruments in both traits.

Table 1: An overview of the methods considered in our comparisons.

The simulations corresponding to the LCV null model are defined by 1) an equal effect of the hidden confounder on both observed traits (i.e., $\eta = \nu$) and 2) a genetic architecture of U that results in an equal per-variance contribution of each SNP to A and B both when it acts directly on them directly or through U. As in [5], we evaluated settings where the studies for A and B had 1) equal power ($N_A = N_B = 100,000$) and an equal genetic architecture, 2) equal sample sizes but trait B was less polygenic, 3) study B had reduced power ($N_B = 20,000$) and 4) study B had reduced power while being less polygenic. In each setting, the heritability of all phenotypes was 30% ($h^2_A = h^2_B = h^2_U = 0.3$). In all settings we simulated $500,000$ total SNPs. With equal genetic architecture, both observed traits had $2,500$ total (direct plus shared) causal SNPs. When trait B had lower polygenicity, it had half as many directly causal SNPs. For example, when U has 500 causal SNPs, A has an additional $2,666$ direct SNPs while B has $1,333$ additional direct SNPs. We varied ν, η and the proportion of shared causal SNPs such that the induced genetic correlation varied from 0.0 to 0.6. For complete settings for each simulation see Table S3.

The simulations corresponding to the CAUSE model are defined by 1) a stronger effect of U on the exposure relative to the outcome (e.g. $\frac{\nu}{\eta} = \sqrt{0.05}$) and 2) a genetic architecture of the hidden trait that results in an equal per-variance contribution to the A whether the SNP acts directly on it or via U. Like [3], we chose four broad categories wherein we adjusted the proportion of the causal variants acting through U, q, from 0.0 to 0.33. The four categories correspond to 1) equal power ($N_A = N_B = 100,000$) with a stronger shared effect ($\frac{\nu}{\eta} = \sqrt{0.05}$), 2) equal power with a weaker shared effect ($\frac{\nu}{\eta} = \sqrt{0.02}$), 3) study 2 having lower power ($N_B = 20,000$) with the stronger shared effect and 4) study 1 having lower power ($N_A = 20,000$) with the stronger shared effect. In these settings the heritability of the observed phenotypes was 25% ($h^2_A = h^2_b = 0.25$) with $1,000$ total (direct plus shared) causal SNPs per observed phenotype out of $500,000$ SNPs. We set $\nu = \sqrt{h^2_A}$ so that the heritability of the latent factored varied from 0 to 50% as we varied q. For complete settings for each simulation see Table S4.

We exhaustively tested all combinations of 1) low (500 directly causal variants) and high ($2,000$ directly causal variants) polygenicity for each of A, B, and U, 2) either equal ($100,000$ individuals per study) or unequal ($25,000$ individuals in the under-powered study) sample sizes, and 3) either equal effects of U on both traits ($\eta = \nu = \sqrt{0.3}$) or unequal effects ($\eta = \sqrt{0.3}, \nu = \sqrt{0.1}$ or $\eta = \sqrt{0.1}, \nu = \sqrt{0.3}$). For complete simulation settings see Table S5.

We assessed the power of WWER under the one and two-directional alternative hypothesis with no correlated pleiotropy as compared to other methods. In these settings $\eta = \nu = 0$ and one or both of γ and δ are > 0. In our simulations under the unidirectional alternative, we varied the proportion of variance in B explained by A from 1% ($\gamma = \sqrt{0.01}$) to 20% ($\gamma = \sqrt{0.2}$). In all settings, the heritability of both phenotypes was 25%. We exhaustively tested all combinations of low and high polygenicity (500 or $2,000$ directly causal variants, respectively) as well as power ($N = 100,000$ in the high powered study, $N = 25,000$ individuals in the low powered study). For complete simulation settings see Table S12. In our simulations under the bidirectional alternative we evaluated power to detect both effects ($A \to B$ and $B \to A$) simultaneously. We set δ to either $\sqrt{0.01}, \sqrt{0.03}, \sqrt{0.10}$ and varied γ from $-\sqrt{0.1}$ to $\sqrt{0.1}$ and again adjusted the polygenicity of each trait and the sample size of each study. For complete simulation settings see Table S14.

Selection of phenotypes for analysis

We obtained summary statistics for sex-split UK Biobank phenotypes from the Neale lab, who corrected for age, age^2 and 20 principal components of the genotype matrix [15]. For ease of interpretation, we transformed all effect sizes to the per-variance scale. We filtered for phenotypes with an LD score regression heritability Z-score above 4 and removed phenotypes defined as "low" confidence, defined by the Neale lab as effective sample size $< 20,000$, standard error $> 12\times$ the expected standard error given sample size, or bad ordinal coding. We also removed one phenotype from every pair with genetic correlation above 0.9 to avoid including what are effectively duplicate traits. We used male summary statistics for SNP selection and weight estimation, and female summary statistics for effect estimation. We removed any trait with an estimated male-female genetic correlation $\rho_g < 0.5$ or a Z-score for non-zero genetic correlation below 2. We used LD-pruned SNPs attaining genome-wide significance ($p \leq 5 \times 10^{-8}$) as instruments with WWER to estimate causal effects (CE). We first consider measurements of blood composition, biomarkers and IMIDs. After filtering, we had 21 measurements of blood composition, 20 blood biomarkers and 10 IMIDs (Table S20). Next, we consider a broader analysis of the entire UKBB dataset. After filtering, we had 411 total phenotypes (Table S23). Of the 411 phenotypes chosen for analysis, 153 had at least 5 independent GWAS significant loci (Table S23). We used WWER to estimate the CE of all 153 phenotypes with at least 5 GWAS-significant loci on all 411 phenotypes. This results in bi-directed effect estimates for the $11,628$ pairs of traits where both have at least 5 instruments, and uni-directional effect estimates for the remaining $39,474$ pairs for a total of in $62,730$ CE estimates. Because this analysis constitutes a large number of additional hypotheses, all q-values reported in this manuscript, including those focused only on the biomarker and IMID analysis, have been corrected using the Benjamini-Hochberg technique for all $62,730$ tested hypotheses.

3 Results

Simulations

WWER reduces false discoveries due to correlated pleiotropy

Our first goal was to assess the calibration of WWER under the two-way null as compared to other methods under a broad range of simulation settings (Figure 2). Our first set of simulations was designed to mirror those in the LCV study [5], varying the power of each study and genetic architecture of the observed phenotypes under a range of genetic correlation values (for details see Section 2). WWER and Egger regression with Stieger filtering maintained a low false positive rate across all of these simulation settings, while other methods had variable performance depending on the setting (Figure 2A). CAUSE also performed well unless the genetic correlation was 0.4 or above and the studies had unequal power, while MBE and MRMix also struggled with higher values of genetic correlation. Egger regression, MR PRESSO, IVW and raps all performed poorly overall. For space considerations we have omitted raps and WME from Figure 2, for complete results see Tables S6-S7.

Our second set of simulations was designed to mirror those in the CAUSE study [3]. We varied the strength of the effect of the shared variable on the outcome and the power of each study for a range of proportions of shared variants (for details see Section 2). In this setting, CAUSE, all Egger-based methods as well as MRMix and the MBE perform similarly well with a well-controlled error rate at lower proportions of shared variants and some excess false positives at higher levels. The WME, MR PRESSO, and (r)aps perform similarly to IVW which struggles to control false positives even for relatively small fraction of pleiotropic SNPs in all settings. CAUSE seems to perform better here than in similar situations in the original manuscript [3]. This is likely because we are using pre-pruned variants without linkage disequilibrium (LD), unlike in [3] where CAUSE must additionally handle LD. In the B to A direction, all methods are able to control the false positive rate. For complete results see Tables S8-S9.

In the aforementioned simulations, the genetic architecture of the hidden node is implicitly tied to those of the observed variables (see Section 2 and Supplemental Methods for more details). In our final set of simulations we sought to decouple these architectures and explicitly manipulate the genetic architecture of the hidden node. We exhaustively tested all combinations of high and low polygenicity for each trait while varying the power of each study and the effect of the hidden factor on the observed traits (for details see Section 2). Due to space limitations we present 8 of the 42 resulting combinations in Figure 2C and the rest in Tables S10-S11. Perhaps unsurprisingly, some settings favor certain methods over others and there is no method that controls the false positive rate (FPR) in every setting. WWER and Egger regression with Steiger filtering worked well for most settings. However, there were notable settings where these methods performed poorly. For example, when the polygenicity of A is high, the polygenicity of U is low, U has a larger effect on A, and the sample size of A is large, both methods produce a false positive $> 95\%$ of the time. (Table S10 lines 5 and 33, Table S11 lines 2 and 16). Another interesting setting is low polygenicity of both A and B, high polygenicity of U, equal sample sizes and a larger effect of U on the exposure (Table S11 line 14). In this setting WWER produces many excess false positives ($FPR = 0.47 \pm 0.05$), but this is roughly half as many as standard Egger regression or Egger with Steiger filtering ($FPR = 0.88 \pm 0.03$ and $FPR = 0.80 \pm 0.04$, respectively). With some exceptions, CAUSE and MBE also performed well in these settings, while all other methods performed poorly overall.

We summarize our results in Table 2 by ranking the methods according to FPR and mean absolute error (MAE) in each of the simulations settings in both directions. For example, in a given simulation setting the method with the lowest FPR receives a rank of 1, the next lowest receives a rank of 2, and so on. We then calculate the mean ranking across all simulation scenarios for each method. We also calculate the percentage of simulation settings where each method had an estimated FPR whose 95% confidence interval contained either 0.05, indicating well-calibrated p-values, or 0.20, indicating that the excess false positives are limited to a useful level. A perfectly-calibrated method would have an estimated FPR at or below 5%

Method	FPR rank	MAE rank	FPR < 5%	FPR < 20%	Runtime (s)
WWER	6.034	6.735	77.4	92.7	0.634
Steiger	5.329	7.372	76.2	92.1	0.634
CAUSE	4.210	4.671	81.7	90.9	2958.892
MBE	3.351	5.823	84.8	89.6	38.140
MRMix	4.125	6.134	76.8	87.2	79.670
Egger	6.683	9.183	59.8	67.1	0.632
WME	6.372	5.177	52.4	63.4	8.594
MR PRESSO	8.287	4.366	45.7	54.3	313.527
raps	8.381	6.421	39.0	50.0	0.684
IVW	9.418	6.683	36.0	40.2	0.640
aps	9.290	8.116	35.4	39.6	0.635

Table 2: A summary of the results from all of our null simulations. In each setting, we ranked every method according to its false positive rate (FPR) and mean absolute error (MAE). Then, we calculated the mean ranking of each method across all simulations settings (columns FPR rank and MAE rank, respectively). We also calculated the percentage of settings in which each method had $FPR < 5\%$, indicating well-calibrated p-values, as well as the percentage of settings with $FPR < 20\%$, indicating a controlled level of excess false positives (columns FPR < 5% and FPR < 20%, respectively. Finally, we calculated the time required to calculate an effect in each direction with each method (column Runtime).

in every simulation setting and receive a score of 100. However, some simulation settings are particularly challenging, for example when the hidden node explains $> 50\%$ of the variance in the observed traits, or when $> 50\%$ of the instruments for the exposure are shared with the latent variable. The MBE, CAUSE, and MRMix performed quite well overall. These three methods generally produced the best-calibrated p-values and controlled the FPR at 5% in the highest percentage of tested cases. However, WWER followed by Egger with Stieger filtering produced a controlled amount of excess false positives ($FPR < 20\%$) more frequently than other methods. WWER generally produced slightly less conservative p-values while having a lower MAE overall as compared to Egger with Stieger filtering. While these two methods did perform similarly, as mentioned above we found evidence of cases where WWER out-performed Egger with Stieger filtering by a large margin, but the opposite was never true (Tables S8-S9). A final consideration, especially in exploratory data analysis applications, is run-time. Regression-based methods are very fast, while more sophisticated methods can take much longer. CAUSE took nearly 50 minutes on average to calculate effects in both directions (Table 2). While the MBE and MRMix are somewhat faster, we used only a small number of sampling iterations (1000) to generate p-values accurate to two significant digits. In exploratory data analysis cases, where the multiple testing burden is likely to be high, many more iterations will need to be used to generate p-values with more significant digits.

WWER maintains power

Our next goal was to evaluate the power of WWER as compared to other methods when the alternative hypothesis is true and there is no correlated pleiotropy. We varied the strength of the causal effect, polygenicity of each trait, and the sample size of each study (see Section 2). WWER and the other regression methods Egger regression and Stieger filtering performed similarly well, with generally strong performance for effect sizes above $\sqrt{0.01}$ but reduced performance when A was highly polygenic but its study was under-powered. MR PRESSO and CAUSE show generally improved power in these more difficult cases, while the MBE can improve power when the study for A is under-powered but reduce it when the study for B is under-powered. MRMix generally performs poorly compared to the other methods. Complete results are given in Table S15.

Next, we considered the bi-directional alternative hypothesis, where both phenotypes have a causal effect on each other. We tested several combinations of joint effect sizes while again varying the genetic architecture

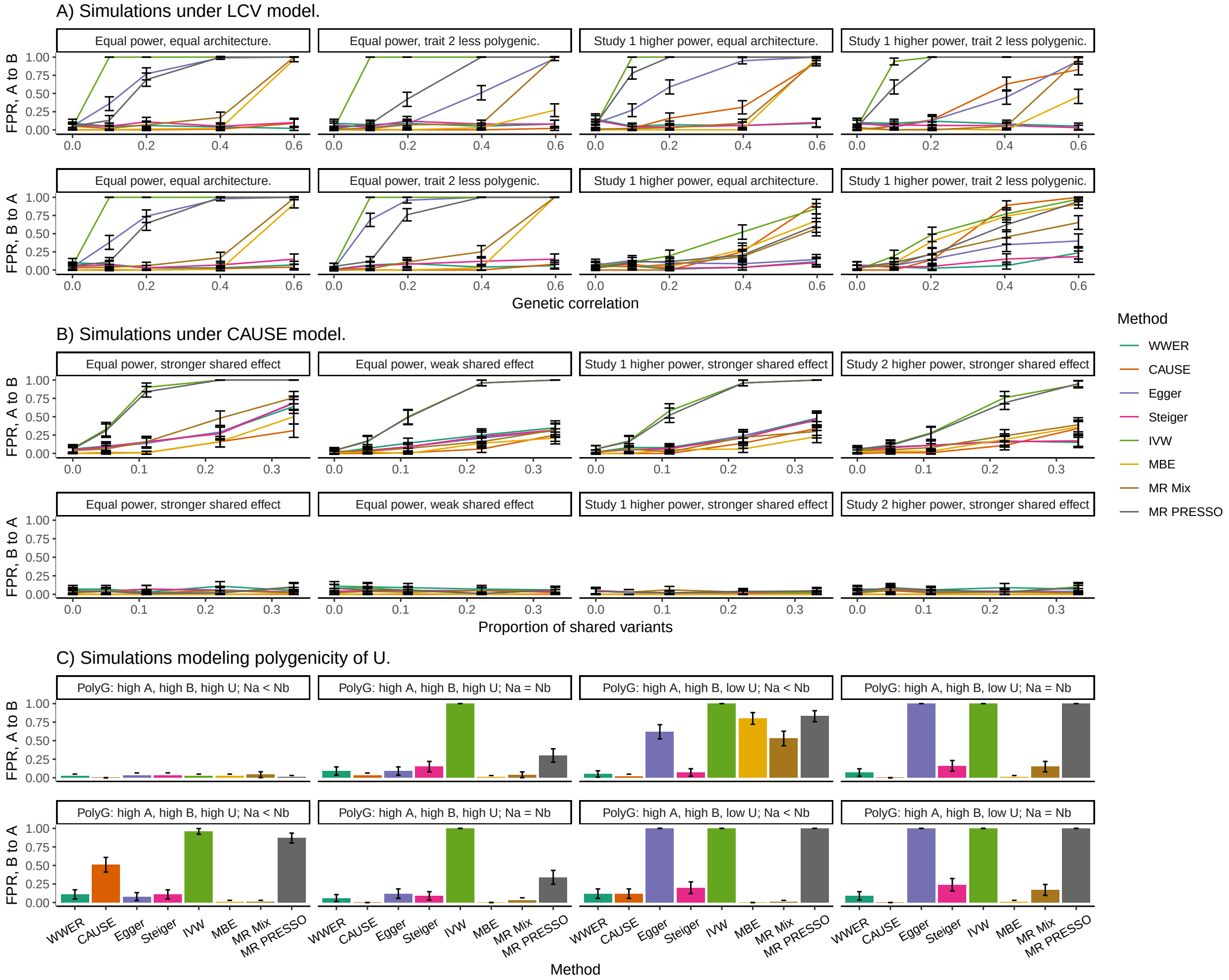


Figure 2: False positive rate in simulations under the bi-directional null for various settings of the simulation parameters. In all cases we consider both the A to B direction (top) and B to A direction (bottom). a) Simulations with parameters set to mimic the LCV model while varying the power and polygenicity of each trait (panels) as well as the genetic correlation (x-axis). WWER and Steiger filtering perform well, while other methods struggle with at least one setting. b) Simulations with parameters set to mimic the CAUSE model while varying the power and strength of the effect of the hidden node (U) on the observed traits A and B (panels). All methods with the exception of IVW and MR PRESSO perform well. c) Simulations explicitly modeling the polygenicity of A, B and U, while varying the relative power of each study (panels). In the panels shown, there is a strong symmetric effect of the hidden node on the traits. Simulations with asymmetric effects are shown in Tables S8-9. There is no method that performs well in every setting, but WWER, Stieger filtering, CAUSE, the MBE, and MRMix perform well overall. Our results are further summarized in Table 2.

and power of each study (see Section 2 for details). Many trends from the unidirectional alternative were replicated here. Specifically, CAUSE, IVW and MR PRESSO performed well overall. The regression based methods performed similarly well in most settings, with lower power when the sample-sizes were unequal. The MBE was generally out-performed by regression based methods when the studies had equal sample size, but the opposite was true for unequal sample sizes, especially when the effects were larger. MRMix again had poor power overall. Interestingly, we found two settings where standard Egger regression was substantially out-performed by both WWER and Egger with Steiger filtering: $\delta = \sqrt{0.2}$, high polygenicity of A, low polygenicity of B, and either equal sample sizes or a larger sample size for A (Figure 3D-F). Complete results are given in Table S19.

Since we are concerned with estimating effects in both directions, we must take care to verify that under the unidirectional alternative, high power in the A to B direction (alternative hypothesis is true) does not result in a high false positive rate in the B to A direction (null hypothesis is true). In Figure S1 we plot the FPR in the B to A direction as a function of γ for each corresponding simulation in Figure 3A-C. All methods except IVW, standard Egger regression, and MR PRESSO were able to control the reverse false positive rate. This was primarily an issue for larger effects ($\gamma = \sqrt{0.1}$ and $\gamma = \sqrt{0.2}$) when study B had high power. Complete results are given in Table S16.

Finally, we considered the one-way alternative hypothesis in the presence of correlated pleiotropy ($\gamma, \eta, \nu > 0$). In this setup both studies had equal power and we varied the strength and symmetry of the effect of U, considering equal weak pleiotropy ($\eta = \nu = \sqrt{0.02}$), equal stronger pleiotropy ($\eta = \nu = \sqrt{0.05}$), or unequal pleiotropy with a stronger effect on either A or B. For complete simulation parameters see Table S13. On the one hand, pleiotropy increases the power to detect the non-null effect because it lends additional signal supporting the effect of A on B (Figure S2). On the other hand, it also leads to additional false positives in the reverse direction (Figure S3). Here MR PRESSO and IVW produce a false positive nearly all the time if there is a strong pleiotropic effect on A. Standard Egger regression performs worse here than in the simpler setting of a true alternative hypothesis with no pleiotropy, but WWER and Stieger filtering are able to reduce this false positive rate substantially. Here, WWER clearly out-performs Steiger filtering in settings where they both produce excess false positives, such as when the polygenicity of A and B are high but the polygenicity of the confounder is low. For complete results see Table S17-18.

Comparison within another simulation framework

Recently, Qi and Chatterjee [16] conducted a thorough simulation study comparing 10 MR methods using realistic simulation settings. We were curious to contextualize the performance of our method within these known results. Therefore, we modified their software to additionally run WWER, and used it to evaluate performance in an additional 40 simulation scenarios. Their comparisons include many of the methods discussed here, but also include some that we have not investigated such as the Contamination Mixture [17] and Robust MR [18]. For descriptions and assumptions of these additional methods, see Table S1. Their simulation approach is similar to ours in that they examine realistic settings for the heritability of the trait, proportion of genome-wide causal variants, proportion of the variance in the exposure explained by detected instruments, and proportion of the variance in each phenotype explained by correlated pleiotropy. The models also differ in some respects. For example, [16] use a smaller number of total variants ($200,000$ vs our $500,000$) and explore larger sample sizes (up to $N_A = 1,000,000$). However, the largest difference in these models is their treatment of uncorrelated pleiotropy. In our model, we allow for uncorrelated pleiotropy by sampling effect SNPs independently for each phenotype (A, B, and U), allowing for a proportion of SNPs to randomly effect both A and B or all 3 phenotypes, resulting in both correlated and uncorrelated pleiotropy. In [16], the model assumes that all invalid instruments exhibit both correlated and uncorrelated pleiotropy. We term these variants "doubly pleiotropic" and describe them as having "dual pleiotropy". We were surprised to see that if all invalid variants are doubly pleiotropic, there was a detrimental effect on the performance of WWER at very large sample sizes. For example, with 30% invalid instruments, and $N_A = 200,000$ samples, WWER produced a false positive 24% of the time, while MRMix and MBE were able to control the FPR at the desired level (Figure S4A). With $500,000$ samples the error rate increased to 77% while MRMix and MBE continued to successfully control the FPR.

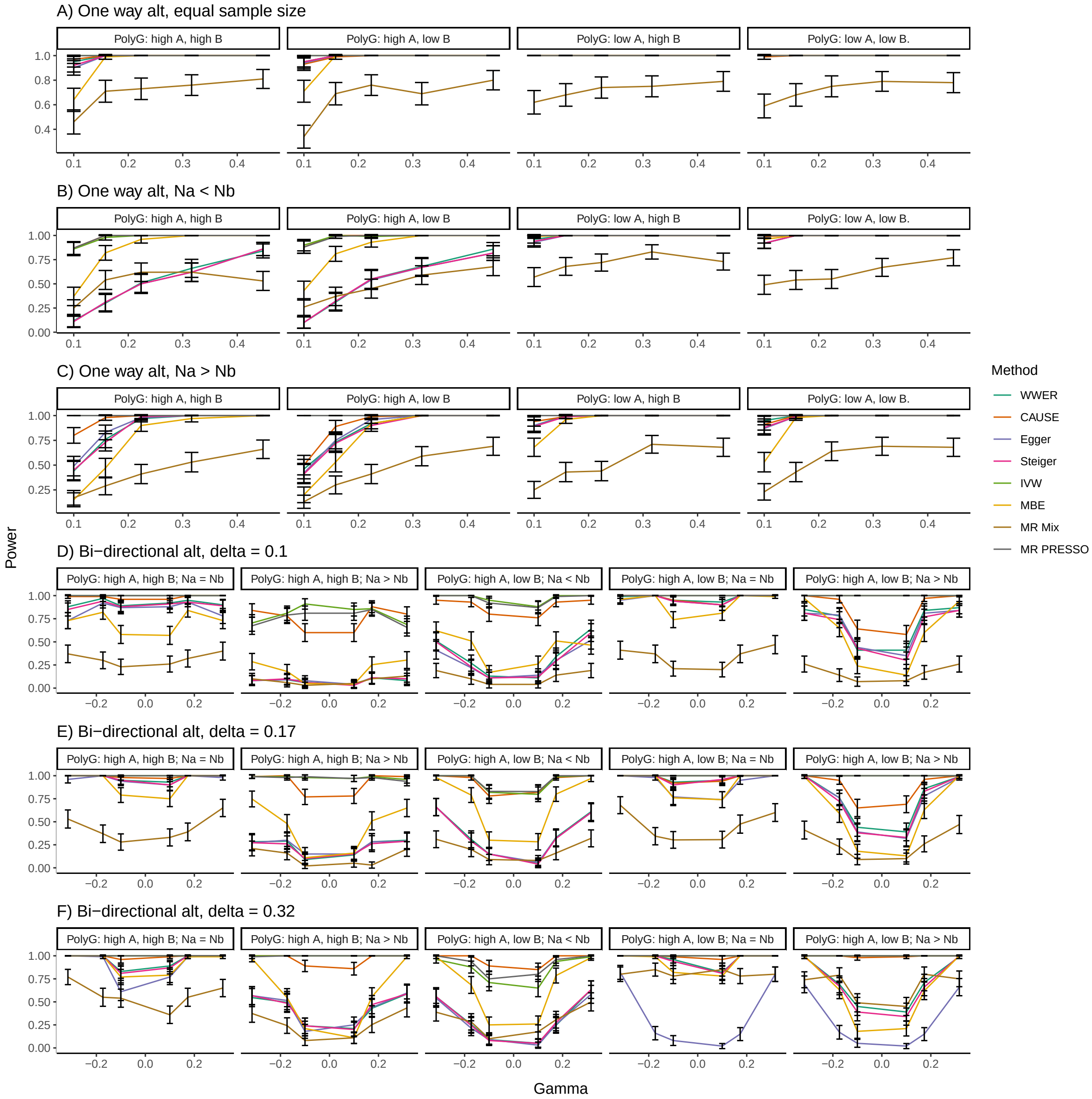


Figure 3: Power analysis of simulations under both the one way (A causes B, a-c) and bidirectional (A causes B and B causes A, d-f) alternative, without additional pleiotropy. a) With equal sample sizes, all methods except MRMix show high power for all settings of the polygenicity of A and B (panels). b) When study A has lower power and the polygenicity of A is higher, regression-based methods have reduced power and are out-performed by the MBE. c) When study A has higher power, the opposite is true. (d-f) The power to detect an effect in both directions for all combinations of polygenicity and power as a function of the effect of A on B for various values of the effect of B on A.

We therefore modified the simulation code provided by [16] to accept an additional parameter governing the proportion of invalid instruments that exhibit both kinds of pleiotropy, and re-ran the simulations such that either 50% (Figure S4B) or 0% (Figure S4C) of variants were doubly pleiotropic. In both of these settings, WWER was able to maintain a low FPR. Therefore, we conclude that WWER only suffers from increased false positives when the vast majority of variants exhibit dual pleiotropy. We also used their simulation framework to evaluate the effect of directional (non-zero mean) pleiotropy, with the proportion of doubly pleiotropic variants set to 50%. We found the performance of WWER did not substantially differ from the unbalanced setting (Figure S4D). Finally, we considered the power under the one-way alternate hypothesis with 30% of variants exhibiting uncorrelated pleiotropy. We found the power of WWER exceeded Egger regression, but generally fell short of MBE, MRMix and the other methods (Figure S5).

Qi and Chatterjee also suggest that a downsampling technique can be used to diagnose effects of problematic pleiotropy [16]. They observe that methods that produce false positives tend to have increasing estimates of the effect size as a function of sample size, while well-behaved methods do not. Therefore, we plotted the mean effect estimate with mean standard error estimate vs sample size under the null with 50% invalid instruments as we varied the proportion of doubly pleiotropic variants from 1 (the default) to 0.5 and 0 (Figure S6). We found that WWER only showed increasing effect estimates with sample size when all variants were doubly pleiotropic. We conclude that in secondary analysis of specific phenotypes, downsampling could potentially be used to diagnose cases where WWER produced false positives.

UK Biobank Analysis

Application to blood traits and immune disorders

There are a number of common disorders involving immune system and inflammatory response disregulation (immune mediated inflammatory disease, IMID), such as allergy, asthma, diabetes and psoriasis, among others [19]. Blood is both an easily accessible tissue and a heterogeneous mixture of numerous cell types with relevance to inflammatory and immune response, so there is a strong interest in intermediate blood biomarkers of IMIDs for measuring disease risk, monitoring progression, and developing treatments [19, 20]. The UK Biobank (UKBB) contains measurements of clinical laboratory biomarkers, as well as blood cell-type composition and disease phenotype data for $> 480,000$ individuals [21]. We filtered UKBB phenotypes as described in Section 2, leaving 21 measurements of blood composition, 20 blood biomarkers and 10 IMIDs (Table S20). We used LD-pruned SNPs attaining genome-wide significance ($p \leq 5 \times 10^{-8}$) as instruments with WWER to estimate causal effects (CE) of each biomarker on each disease, and vice-versa (disease on biomarker). We found 83 (of 410) significant effects at FDR 5% in the biomarker to disease direction (Figure 4A, Table S21). In the following, we denote adjusted p-values with q. We observed a strong effect of platelet traits on asthma and allergy. For example, increased platelet distribution width (PDW) decreases asthma risk ($CE = -0.034, q = 4 \times 10^{-10}$) and allergy risk ($CE = -0.016, q = 2 \times 10^{-2}$), increased mean platelet volume (MPV) decreases asthma risk($CE = -0.014, q = 2 \times 10^{-2}$) and increased platelet-crit decreases allergy risk ($CE = -0.066, q = 3 \times 10^{-10}$). Platelet traits have long been implicated in asthma and allergy [22, 23, 24], with lower MPV values observed in individuals with asthma and allergy and lower PDW values observed in individuals with asthma. Platelet traits are now thought to play an important role in both the innate and adaptive immune response [25]. We find that PDW is implicated in 7 of the 10 IMID studied, and MPV is implicated in 4 of the 10. This gives evidence that platelet activity can have an effect on immune-system function, with broad downstream consequences that include many common diseases.

Lymphocyte count, a marker of inflammation, is also implicated in 7 of the 10 IMID that we analyze. We detect effects of increased lymphocyte count on psoriasis ($CE = 0.159, q = 1 \times 10^{-9}$), Crohn's disease ($CE = 0.057, q = 3 \times 10^{-5}$), and ulcerative colitis ($CE = 0.037, q = 4 \times 10^{-2}$). We detect effects of decreased lymphocyte count on asthma ($CE = -0.12, q = 6 \times 10^{-9}$), osteoarthritis ($CE = -0.069, q = 4 \times 10^{-6}$), allergy ($CE = -0.101, q = 3 \times 10^{-5}$) and diabetes ($CE = -0.04, q = 3 \times 10^{-3}$). A lower neutrophil to lymphocyte ratio has been observed in patients with each of these diseases [26, 27, 28, 29]. In several of these results, our estimated CE and the genome-wide genetic correlation have opposite signs. For example, the genetic correlation between lymphocyte count and asthma is positive ($\rho_g = 0.054$), as is the genetic correlation between lymphocyte count and osteoarthritis ($\rho_g = 0.212$). In each of these cases, the negative effect direction inferred by WWER is more consistent with the observed lower neutrophil to lymphocyte ratio in these diseases. This indicates that the total genetic correlation can be misleading, even in the presence of a causal effect, it is possible for a genetic confounder, or possibly random noise, to result in an observed genetic correlation with a different sign than the true causal effect.

Total cholesterol also has several disease consequences. We observe protective effects of increased total cholesterol level on diabetes ($CE = -0.047, q = 4 \times 10^{-10}$), deep vein thrombosis (DVT, $CE = -0.035, q = 3 \times 10^{-8}$), diverticulitis ($CE = -0.025, q = 5 \times 10^{-4}$), and emphysema ($CE = -0.016, q = 4 \times 10^{-2}$). We also observe a protective effect of increased HDL cholesterol level on asthma ($CE = -0.026, q = 2 \times 10^{-3}$) These findings are particularly interesting in light of recent work suggesting that cholesterol can lower inflammation [30], that higher cholesterol is a consequence of the body's attempt to control inflammation, rather than the cause of disease in itself [31]. Interestingly we observe a weak effect of increased cholesterol on allergy risk($CE = 0.021, q = 4 \times 10^{-2}$) which is inconsistent with the genetic correlation between these traits ($\rho_g = -0.014$). Cholesterol is known to effect development of allergy, but reports differ on the direction of the effect [32].

Other notable effects we observe include a strong effect of eosinophil percentage on asthma ($CE = 0.118, q = 5 \times 10^{-6}$), aspartate aminotransferase on ulcerative colitis ($CE = 0.047, q = 5 \times 10^{-5}$), glucose on emphysema ($CE = 0.051, q = 9 \times 10^{-5}$), and a protective effect of vitamin D on diabetes ($CE = -0.024, q = 5 \times 10^{-2}$). Eosinophils are known to play an important role in the pathogenesis of asthma [33], with well-established genetic evidence indicating a protective effect of lower eosinophil count on asthma risk [34]. Liver test abnormalities are frequently observed in patients with inflammatory bowl diseases [35] and appear to be an risk factor for complications in patients with Crohn's disease [36]. Blood glucose has been observed to be elevated in patients experiencing chronic obstructive pulmonary disease (COPD) exacerbations [37]. Vitamin D has been linked to the onset of diabetes [38].

We found 36 (of 164) significant effects in the disease to biomarker direction after accounting for multiple testing using the BH procedure (Figure 4B, Table S22). Most of these are driven by just two phenotypes: 20 are effects of psoriasis and 11 are effects of asthma. Some of the top effects of psoriasis are related to red blood cells (RBC). We estimate that psoriasis decreases mean sphered cell volume ($CE = -0.12, q = 1 \times 10^{-7}$), mean corpuscular volume ($CE = -0.15, q = 1 \times 10^{-7}$), and mean reticulocyte volume ($CE = -0.12, q = 4 \times 10^{-5}$), while increasing red blood cell count ($CE = 0.12, q = 1 \times 10^{-7}$). There is an established relationship between red blood cell function and psoriasis [39, 40, 41]. There is disagreement in the literature about the correlation of red blood cell count and psoriasis, with one study showing an increase in patients, consistent with our results [41] and others showing a decrease, inconsistent with our results [39, 40]. However, the latter study also shows that treatment of psoriasis can correct RBC damage, which might suggest that psoriasis is the cause, rather than consequence, of RBC damage. We also observe effects of psoriasis on lipid profile. We infer psoriasis increases HDL cholesterol ($CE = 0.088, q = 3 \times 10^{-4}$), total cholesterol ($CE = 0.097, q = 2 \times 10^{-3}$), and triglycerides ($CE = 0.123, q = 4 \times 10^{-7}$). Psoriasis is well-known to be co-morbid with cardiovascular disease, and dislipidemia has long been observed in patients with psoriasis [42, 43]. However, we note that our inferred direction of effect for HDL cholesterol is inconsistent with prior literature showing decreases in HDL cholesterol level in patients with psoriasis, and with the genetic correlation between the traits ($\rho_g = -0.21$). Psoriasis is known to have a complex effect on HDL cholesterol function [44], and it is likely that the genetic instruments we use to estimate this effect on serum HDL levels do not reflect the complexity of this interaction.

Inferred effects of asthma include decreases in IGF-1 ($CE = -0.308, q = 1 \times 10^{-3}$), lymphocyte percentage ($CE = -0.354, q = 2 \times 10^{-3}$), and monocyte percentage ($CE = -0.205, q = 5 \times 10^{-2}$), and increases in C-reactive protein ($CE = 0.226, q = 2 \times 10^{-2}$) and glycated haemoglobin ($CE = 0.233, q = 2 \times 10^{-2}$). Glycated haemoglobin and C-reactive protein have both been observed to be elevated in patients with asthma [45, 46]. Monocytes and lymphocytes are both known to play an important role in asthma [47, 48, 49], however it is unclear how the impact of recruitment of specific monoctye and lymphocyte subsets to the lungs in asthma patients would impact circulating blood levels of these broad cell types. IGF-1 is known to play a function in the repair of lung tissue [50]. Serum IGF-1 level is known to be anti-correlated with asthma incidence and severity in the UK Biobank [51]. Our results suggest this is a consequence rather than

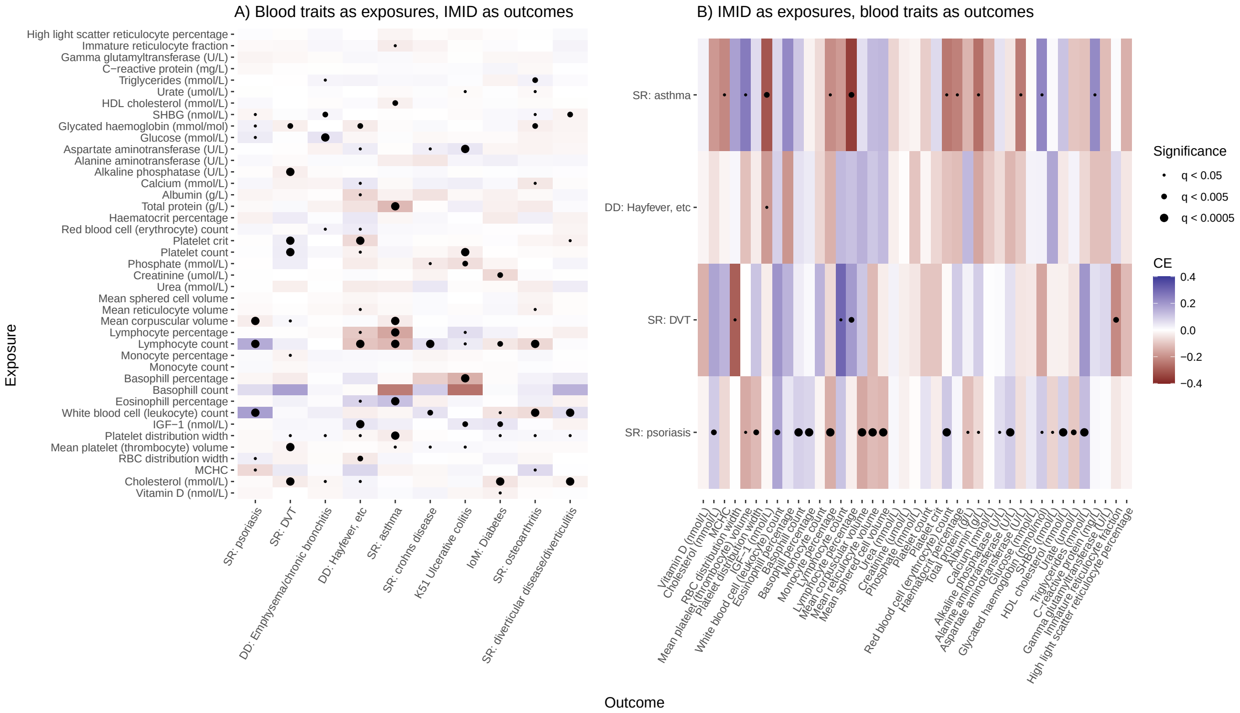


Figure 4: An investigation into the relationship between immune-mediated inflammatory diseases and blood biomarkers in the UK Biobank. A) Estimated causal effects using blood traits as exposures and IMID as outcomes replicates known disease biology. B) Estimated causal effect using IMID as exposures and blood traits as outcomes reveals many significant "reverse" causal effects. Dots indicate level of statistical significance of $p < 0.05$ after FDR correction.

a cause of asthma.

As we have observed certain simulation scenarios where WWER is likely to produce a false positive, for example when a large majority of invalid instruments are doubly pleiotropic, we sought to assess whether this may have had an impact on our results. Strictly-speaking, we cannot apply the aforementioned downsampling approach without using the individual-level data. Instead, a reviewer suggested that we leverage the intuition that very significant associations tend to remain significant at lower sample sizes. We used an increasing sequence of p-value cutoffs ($p = 5 \times 10^{-28}$, 5×10^{-20}, 5×10^{-14}, 5×10^{-10}, 5×10^{-8}) to select instruments and compute weights, then calculated the effect estimate and standard error using each set of instruments. In Figure S10, we plot the effect estimate and 95% confidence interval as we vary the SNP inclusion threshold for every pair of phenotypes mentioned in this analysis. We found two pairs that could potentially be considered problematic. First, the absolute effect of asthma on monocyte percentage increases from 0.146 ± 0.17 to 0.205 ± 0.07. This is a large increase, however the increase is not strictly linear with decreasing threshold, and the standard error of the estimate is high. The other potentially problematic effect is that of psoriasis on cholesterol, which increases from 0.067 ± 0.035 to 0.097 ± 0.024.

Comparison to other methods

To compare the results obtained by WWER to those obtained by other methods, we repeated the blood biomarker and IMID analysis using four additional MR methods: IVW, Egger regression, the MBE, and MRMix. IVW was chosen as a baseline which does not control for pleiotropy, Egger was chosen because it is ubiquitously used and only accounts for uncorrelated pleiotropy, while MBE and MRMix were chosen because they performed well overall in our simulations and were computationally tractable to run in this smaller analysis. We calculated the number of discoveries at a family-wise error rate (FWER) of 5% for each method accounting for 574 tests. WWER had the most discoveries with 51 total, followed by 44 for IVW, 38 for MRMix, 28 for MBE and just 8 for Egger regression. The overlap in discoveries across methods is shown in Figure S7A. The largest sets were the sets of discoveries unique to WWER (21), unique to MRMix (17) and unique to IVW (12), followed by the set of discoveries found by all methods except for Egger regression (10). Furthermore, we calculated the Jaccard coefficient of the discoveries make by every pair of methods, defined as the size of the intersection divided by the size of the union of two sets (Figure S7B). The largest overall Jaccard index was between IVW and MBE (0.385), followed by IVW and WWER (0.377), MBE and MRmix (0.375); and WWER and MBE (0.274). While there were many discoveries unique to the WWER, IVW and MRMix methods, these discoveries tended to be enriched for association signal in the other methods. For example, discoveries unique to WWER have an average χ^2 test statistic in MRMix of 2.61, while discoveries unique to MRMix have an average χ^2 test statistic in WWER of 3.10. The strongest signal detected by WWER but not other methods is an effect of white blood cell count on psoriasis, which reaches nominal significance using MRMix ($p = 0.011$). The next strongest is the aforementioned effect of PDW on asthma, which also shows strong but not FWER significant signal in MRMix ($p = 0.002$).

We were surprised that WWER produced the most discoveries in this analysis, given that IVW is usually considered the most powerful MR method when pleiotropy is not a concern. Thus, we plotted the effect estimate and 95% confidence interval as a function of p-value cutoff for each of the 51 significant pairs in this analysis (Figure S9). For the most part, effect estimates remained very similar as we varied the inclusion threshold, however there were a handful of notable exceptions. The most obvious is the effect of total protein on asthma, which has an absolute effect estimate of 0.068 ± 0.055 when using a cutoff of $p = 5 \times 10^{-28}$ that increases to 0.124 ± 0.025 at $p = 5 \times 10^{-8}$. Other potentially problematic pairs include the previously mentioned effect of psoriasis on cholesterol, and additionally of psoriasis on PDW where the estimate increases from 0.087 ± 0.043 to 0.124 ± 0.029. In fact, four of the top five pairs in terms of change in effect estimate with increasing inclusion cutoff involve psoriasis as the exposure, and we observe no additional phenotype pairs with large changes in effect estimate. This in itself is interesting, and suggests that the broad causal effects of psoriasis on blood biomarkers that we observe could be driven by shared pathways rather than direct causation.

Phenome-wide analysis

The simplicity and speed of our method allows it to easily scale to phenome-wide analysis. After applying our filtering procedure (see Section 2), we had 411 phenotypes for analysis, of which 153 had at least 5 independent GWAS significant loci (Table S23). We used WWER to estimate the CE of all 153 phenotypes with at least 5 GWAS-significant loci on all 411 phenotypes. This results in bi-directed effect estimates for the $11,628$ pairs of traits where both have at least 5 instruments, and uni-directional effect estimates for the remaining $39,474$ pairs for a total of in $62,730$ CE estimates. Of these, we found $5,770$ effects (9.2%) were significant at a 5% FDR. Complete results for all tested pairs of phenotypes are given in Table S24.

We were curious to compare our CE estimates against estimates of genetic correlation in the same dataset. First, we clustered phenotypes by genetic correlation to determine if the patterns observed are shared in the CE estimates. While there are some similar patterns across the two matrices, the structure in the CE estimates is not as well-defined (Figure 5). Indeed, we find that while the CE estimates and genetic correlation estimates are correlated, that correlation is fairly weak ($r = 0.175 \pm 0.004$). This weak correlation seems to be driven by CE estimates with large standard error. Accordingly, restricting our analysis to CE estimates with standard error below 0.05 yields a much stronger correlation ($r = 0.573 \pm 0.005$). In general we found that more significant CE estimates were more similar to estimates of genetic correlation (Figure S8). As expected, the presence of genetic correlation does not indicate a detectable CE, and the causal effect and the total genetic correlation need not even have the same sign. However, strong CEs do frequently result in a total genetic correlation of similar magnitude.

There were several traits with numerous consequences. The top 5 were white blood cell count (WBCC) with 188 effects, cholesterol with 173 effects, lymphocyte count with 172 effects, sex-hormone binding globulin with

154 effects, and body mass index with 147 effects. The top consequence of higher WBCC was an increase in "nervous feelings" ($CE = 0.12, q = 1 \times 10^{-16}$). WBCC is known to be elevated in individuals with depression and anxiety [52], and could reflect an effect of systemic inflammation on mood. The next strongest effect was a decrease in whole body water mass ($CE = -0.236, q = 1 \times 10^{-16}$). While dehydration is well-known to cause elevated WBCC, our results suggest that the opposite may also be true - higher levels of circulating WBCC could cause the body to retain less water. Two other strong effects of WBCC are on morphology, with an increase in WBCC resulting in a decrease in hip circumference ($CE = -0.179, q = 2 \times 10^{-16}$) and sitting height ($CE = -0.23, q = 9 \times 10^{-16}$). One study of Japanese men found that height and WBCC were inversely correlated, and concluded that this association may result from the presence of inflammation [53].

Interestingly, several of the top consequences of high cholesterol seemed to reflect behavioral changes resulting from common medical advice. For example, we found an effect on increased cholesterol on decreased use of butter ($CE = -0.096, q = 1 \times 10^{-16}$) and increased use of "other spread/margerine" ($CE = 0.09, q = 1 \times 10^{-16}$). We also found increased cholesterol caused a decrease in "salt added to food" ($CE = -0.048, q = 1 \times 10^{-16}$), and an increase in "major dietary changes in the last year" ($CE = 0.069, q = 1 \times 10^{-16}$), indicating high cholesterol results in broad dietary changes. This phenomenon extends to choice of pain medication. We detect a positive effect of high cholesterol on aspirin use ($CE = 0.048, q = 8 \times 10^{-13}$) and a negative effect on ibuprofen use ($CE = -0.026, q = 5 \times 10^{-5}$). This is likely to reflect common medical advice for patients at risk of heart disease to choose aspirin, which has long been thought to reduce risk [54], and avoid ibuprofen, which is thought to reduce the effectiveness of aspirin [55]. We also replicate cholesterol as a known risk factor for heart disease ($CE = 0.086, q = 1 \times 10^{-16}$), which likely also accounts for an observed effect of high cholesterol on earlier "fathers age at death" ($CE = -0.069, q = 1 \times 10^{-16}$).

Several of the top consequences of body mass index (BMI) were also behavioral. For example, we observed a negative effect of BMI on using semi-skim milk ($CE = -0.24, p < 1 \times 10^{-16}$), but a positive effect on using skim milk ($CE = 0.305, p < 1 \times 10^{-16}$). We also observe a positive effect of higher BMI on "major dietary changes in the last year" ($CE = 0.236, p < 1 \times 10^{-16}$). These could again reflect behavioral consequences of common medical advice. Other effects of BMI were on blood biomarkers. For example, we observed an effect of higher BMI on higher C reactive protein (CRP, $CE = 0.353, p < 1 \times 10^{-16}$), lower albumin ($CE = -0.222, p < 1 \times 10^{-16}$), and higher urate ($CE = 0.383, p < 1 \times 10^{-16}$). Higher BMI is well known to cause higher serum urate levels [56], adipose tissue is known to induce low-grade inflammation, which can be measured by elevated CRP levels [57], and BMI is a known risk factor for hypoalbuminemia [58]. We find the known effect of BMI on diabetes ($CE = 0.195, p < 1 \times 10^{-16}$), but also find that BMI has broad effects on health and results in a lower "overall health rating" ($CE = 0.149, p = 7 \times 10^{-8}$).

Finally, we checked whether any of the phenotype pairs mentioned in this analysis showed the trend of increasing effect estimate as we swept the SNP inclusion threshold (Figure S10). We noticed one additional phenotype pair that could be considered problematic - the absolute effect of BMI on albumin increases from 0.147 ± 0.056 to 0.222 ± 0.029. We note that this effect is relatively strong even at the strictest ($p = 5 \times 10^{-28}$) inclusion threshold, but does increase substantially as the threshold is relaxed. All other phenotype pairs showed relatively small fluctuations in effect estimate with increasing inclusion threshold.

4 Discussion

We have introduced a model for bi-directional Mendelian randomization with correlated pleiotropy that allows for flexibility in the specification of the genetic architecture for each trait, as well as a simple method for estimating causal effects called Welch-weighted Egger regression (WWER). We have shown that our method reduces false positives due to correlated pleiotropy compared to traditional methods in a broad range of simulation settings that encompass other recently-proposed models, and is fast enough to be applied at-scale. We first applied WWER to a subset of the UK biobank comprising blood biomarkers and inflammatory disorders, and then more broadly to all heritable phenotypes in the biobank. Our initial analysis reiterated the role of platelet traits in the pathogenesis of asthma and allergy, and found that cholesterol and white blood cell count contribute broadly to inflammatory disease, among other findings. Our broad analysis found thousands of causal effects, many of which stem from a handful of broadly-impactful phenotypes. We replicate

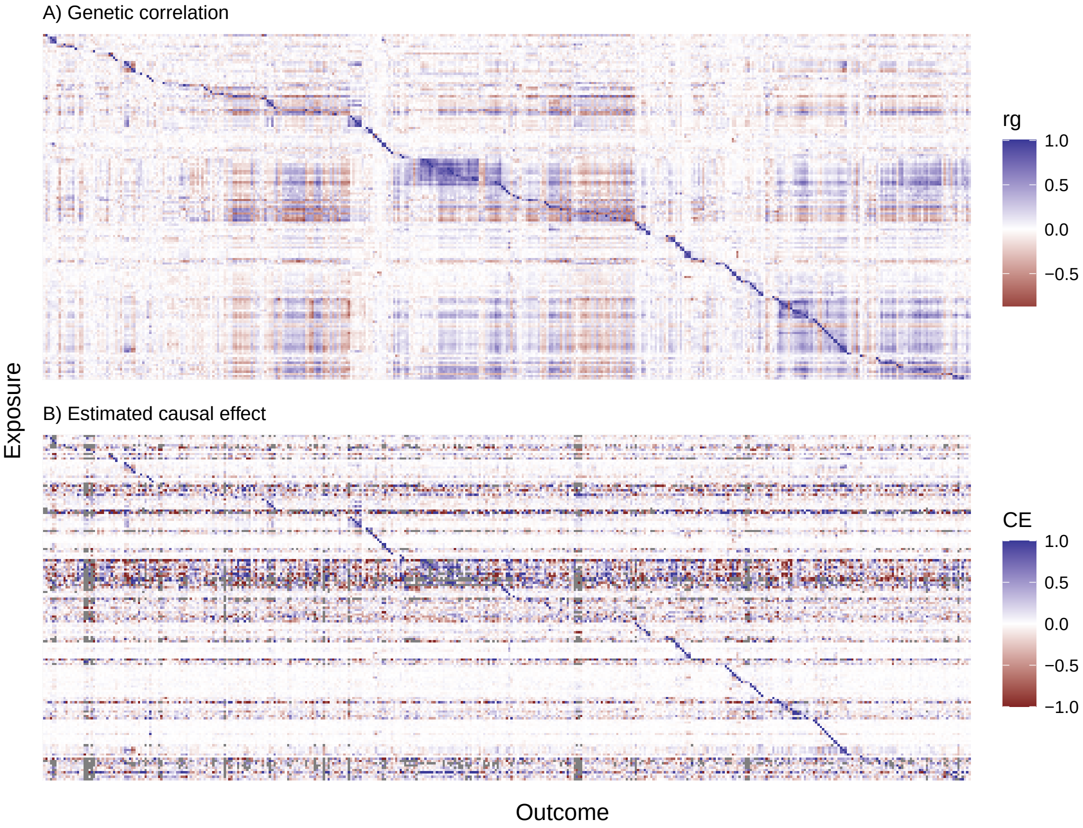


Figure 5: A comparison of genetic correlation (A) with the estimated causal effect (B). We calculated causal effects for all pairs of phenotypes passing our inclusion criteria using LD-pruned GWAS-significant variants as instruments. If both traits did not have significant variants, we calculate a unidirectional effect where the trait with significant variants is the exposure, whereas if both traits have significant variants we calculate a bidirectional estimate. Gray entries in (B) indicate pairs where the exposure had no remaining instruments after filtering likely pleiotropic SNPs, resulting in an NA value. Phenotypes are arranged by clustering on genetic correlation of traits that remain as exposures.

several known risk factors for disease such as high cholesterol on heart disease and high BMI on diabetes, but also detect numerous behavioral changes that seem to result from common physician advice.

Our approach builds on recent MR literature. By filtering genetic variants that have a statistically indistinguishable effect on both the exposure and the outcome, our method is closely related to Steiger filtering [9], which was conceptualized as a method for inferring the effect direction and has not received attention as a method for controlling for correlated pleiotropy. The primary conceptual difference is that we use the test statistic as a regression weight when calculating the effect of the exposure on the outcome with the retained SNPs. Compared to Steiger filtering, we control the FPR in slightly more of the tested settings, while also producing estimates with a lower MAE. There are a small number of settings in which WWER produces a much lower false positive rate than Stieger filtering, but the reverse is never true. However, we find that both methods are generally useful for controlling for correlated pleiotropy. Our approach can be viewed a simple heuristic for classifying variants as effecting the exposure, the outcome, or the hidden variable. More sophisticated mixture model based methods, such as MRMix and CAUSE, are also based on fitting the causal effect using a subset of SNPs that appear to effect the exposure. While these methods also work well in our simulations, they can take a prohibitively long time to run, preventing their application at the scale considered here. By removing genetic instruments with ambiguous effects, our method sometimes filters all potential instruments and cannot estimate the effect. We view this as both an advantage and a disadvantage: we avoid estimating an effect in ambiguous cases, but cannot always produce an estimate.

In our comparisons to other methods using data simulated from our model, we found that there were several settings where WWER had higher power under the alternate hypothesis than some other methods, including MRMix and MBE. However, in our simulations using the model from [16], MRMix and MBE generally had higher power. The primary difference between these scenarios is that the former simulations do not include invalid instruments, while the latter include 30% invalid instruments exhibiting uncorrelated pleiotropy. Indeed, we observed that the presence of some uncorrelated pleiotropy actually improved the performance of MRMix and MBE under both the null and the alt.

In our comparisons against other methods on the UKBB blood biomarker and IMID data, WWER had the most discoveries, followed closely by IVW and MRMix. While we do observe some simulation settings in which WWER has higher power than MRMix, it is unclear why WWER yields more total discoveries than IVW. On the one hand, IVW is generally considered the most powerful approach under the alternate in the absence of pleiotropy, and this is replicated in our simulation studies. On the other hand, WWER similarly produces a lower false positive rate than IVW under the null in every setting we considered. To investigate this further, we performed an approximate downsampling analysis by varying the SNP-inclusion threshold and checking whether effect estimates increased with decreasing threshold. We observed that a small number of our FWER-significant discoveries showed this trend, but these were primarily driven by a single exposure (psoriasis). Moreover, we have calculated effect estimates and standard errors at 5 SNP inclusion thresholds for every phenotype pair considered in this study. While manual inspection of the $5,770$ FDR-5% significant pairs of phenotypes is beyond the scope of this study, we inspected the trend for each of the 53 phenotype pairs mentioned in the main text and found only 3 that might be interpreted as showing a trend representative of pleiotropy. Thus, we have demonstrated that careful secondary analysis can be used to flag potentially-problematic causal effect claims arising from broad initial analyses. We view WWER as a useful technique for exploratory data analysis, where follow-up investigations with other, more computationally-intensive methods or downsampling approaches could be used to increase confidence in specific causal claims.

Because our approach is based on a heuristic, it lacks a rigorous theoretical basis and we cannot make guarantees about settings in which the FPR will be controlled. That said, there are some assumptions that should be satisfied for the method to work as intended. First, the genetic architecture of the instrument discovery and effect estimation cohorts must be identical. In this analysis, we used sex-split summary statistics because they had been made available by the Neale lab, however, we must therefore limit our analysis to phenotypes with high male-female genetic correlation. Next, our method relies on the existence of detectable direct-acting genetic instruments on the exposure. If there are no or very few variants that receive high weights, the weighted regression will be still be dominated by pleiotropic variants and the likelihood of a false positive will increase. Finally, our method performs poorly when the vast majority of variants exhibit both correlated and uncorrelated pleiotropy. While it is likely that some variants exhibiting correlated pleiotropy will also exhibit uncorrelated pleiotropy, it is not obvious that this will be true for all variants in common cases. Moreover, we have thoroughly demonstrated that downsampling can be used to identify cases where the WWER effect estimate increases with sample size, which can flag potentially problematic phenotypes for secondary analysis.

Despite its advantages, our approach has several limitations. First, our method requires that we split the initial cohort into instrument discovery and effect estimation sub-cohorts. While this approach is common in MR methods that must first identify instruments, this reduces power and two sets of summary statistics are not always available. Other recent approaches, such as CAUSE and LCV, have the advantage of modeling the entire spectrum of SNP-trait associations. Second, while our method reduces excess false positives, it does not completely eliminate them. Therefore, a small but notable number of statistically significant results in any large-scale analysis may be due to correlated pleiotropy. We have shown that these failure cases usually correspond to situations where the hidden factor has a strong effect on the exposure, and the exposure does not have many independent large-effect instruments. In this setting, the genetic signature of the exposure and hidden variable are difficult to distinguish. However, the fact that the hidden trait is highly causal for the exposure indicates that these cases may still be biologically interesting, even if they are not directly causal. One advantage of our method is that it only requires GWAS summary statistics, which are both legally and practically easier to share, and faster to work with when the primary data is large [59]. However, summary statistics are inherently limiting. Their use relies on the assumption that the creator of the summary statistics properly controlled for the relevant factors, which may not always be the case when the data are curated by groups without specific expertise in each of the relevant phenotypes. A final limitation of our method is that it estimates the total effect of the exposure on the outcome, which may be mediated by other measured or unmeasured factors.

As biobanks continue to grow in size and scope, new methods that are able to leverage their power while overcoming common pitfalls are required. These datasets offer unprecedented opportunity to study the causal relationship between biomarkers, complex traits and diseases. Broad analysis of the shared genetic effects of pairs of traits can be used to generate causal hypotheses that are much more likely to reflect biologically or medically-relevant phenomena than correlative analyses. It is important to point out that MR analyses without a mechanistic understanding of the biological action of each instrument are inherently speculative, with some researchers suggesting that these instead be called "joint association studies" [60]. This is especially true in large-scale analyses of noisy data, such as population-level biobanks. Indeed, we produce results that are temporally impossible; someone's cholesterol level cannot literally cause their father's heart disease. Nevertheless, the interpretation of this result is clear; many of the risk alleles for cholesterol will be inherited from the father, who will have had more time to develop heart disease, resulting in high power to detect an effect. While a mechanistic understanding of the effects of each genetic instrument is ideal, there is substantial interest in the community in both applying and developing methods for causal inference using statistically associated genetic instruments. We have shown our approach is broadly useful for exploratory analysis of putatively causal effects in ever-growing databases of genotypes and phenotypes.

5 Description of supplemental data

Supplemental data include 10 figures, 24 tables, and supplemental methods.

6 Acknowledgments

BCB and DAK would like to thank Tuuli Lappalainen for helpful feedback on the manuscript. BCB is funded by post-doctoral fellowship from the Data Science Institute at Columbia University.

7 Declaration of interests

The authors declare no competing interests.

8 Web resources

UK Biobank summary statistics are available from the Neale lab at `https://www.nealelab.is/uk-biobank`.

9 Data and code availability

Our full data analysis results are available at `https://zenodo.org/record/4605239`. All code used in the production of this manuscript is available at `https://github.com/brielin/WWER`.

References

[1] Jack Bowden, George Davey Smith, and Stephen Burgess. "Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression". In: *International Journal of Epidemiology* 44.2 (2015), pp. 512–525. ISSN: 14643685. DOI: 10.1093/ije/dyv080.

[2] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. "Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption". In: *International Journal of Epidemiology* 46.6 (2017), pp. 1985–1998. ISSN: 14643685. DOI: 10.1093/ije/dyx102.

[3] Jean Morrison et al. "Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics". In: *Nature Genetics* 52.7 (July 2020), pp. 740–747. ISSN: 15461718. DOI: 10.1038/s41588-020-0631-4. URL: https://doi.org/10.1038/s41588-020-0631-4.

[4] Guanghao Qi and Nilanjan Chatterjee. "Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects". In: *Nature Communications* 10.1 (Dec. 2019), pp. 1–10. ISSN: 20411723. DOI: 10.1038/s41467-019-09432-2. URL: https://doi.org/10.1038/s41467-019-09432-2.

[5] Luke J O'Connor and Alkes L Price. "Distinguishing genetic correlation from causation across 52 diseases and complex traits". In: *Nature Genetics* 50.12 (2018), pp. 1728–1734. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0255-0. URL: https://doi.org/10.1038/s41588-018-0255-0.

[6] N. J. Timpson et al. "C-reactive protein levels and body mass index: Elucidating direction of causation through reciprocal Mendelian randomization". In: *International Journal of Obesity* 35.2 (2011), pp. 300–308. ISSN: 03070565. DOI: 10.1038/ijo.2010.137.

[7] Rebecca C. Richmond et al. "Assessing Causality in the Association between Child Adiposity and Physical Activity Levels: A Mendelian Randomization Analysis". In: *PLoS Medicine* 11.3 (2014). ISSN: 15491676. DOI: 10.1371/journal.pmed.1001618. URL: https://pubmed.ncbi.nlm.nih.gov/24642734/.

[8] Joseph K. Pickrell et al. "Detection and interpretation of shared genetic influences on 42 human traits". In: *Nature Genetics* 48.7 (2016), pp. 709–717. ISSN: 15461718. DOI: 10.1038/ng.3570. URL: http://dx.doi.org/10.1038/ng.3570.

[9] Gibran Hemani et al. "Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome". In: (2017). DOI: 10.1101/173682. URL: https://doi.org/10.1101/173682.

[10] B. L. WELCH. "The generalisation of student's problems when several different population variances are involved." In: *Biometrika* 34.1-2 (1947), pp. 28–35. ISSN: 00063444. DOI: 10.1093/biomet/34.1-2.28. URL: https://pubmed.ncbi.nlm.nih.gov/20287819/.

[11] Marie Verbanck et al. "Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases". In: *Nature Genetics* 50.5 (May 2018), pp. 693–698. ISSN: 15461718. DOI: 10.1038/s41588-018-0099-7. URL: https://doi.org/10.1038/s41588-018-0099-7.

[12] Qingyuan Zhao et al. "Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score". In: *Annals of Statistics* 48.3 (June 2020), pp. 1742–1769. ISSN: 21688966. DOI: 10.1214/19-AOS1866. arXiv: 1801.09652. URL: https://projecteuclid.org/euclid.aos/1594972837.

[13] Jack Bowden et al. "Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator". In: *Genetic Epidemiology* 40.4 (May 2016), pp. 304–314. ISSN: 10982272. DOI: 10.1002/gepi.21965. URL: https://pubmed.ncbi.nlm.nih.gov/27061298/.

[14] Gibran Hemani, Kate Tilling, and George Davey Smith. "Orienting the causal relationship between imprecisely measured traits using GWAS summary data". In: *PLOS Genetics* 13.11 (Nov. 2017). Ed. by Jun Li, e1007081. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007081. URL: https://dx.plos.org/10.1371/journal.pgen.1007081.

[15] *UK Biobank — Neale lab*. URL: http://www.nealelab.is/uk-biobank (visited on 04/09/2021).

[16] Guanghao Qi and Nilanjan Chatterjee. "A comprehensive evaluation of methods for Mendelian randomization using realistic simulations and an analysis of 38 biomarkers for risk of type 2 diabetes". In: *International Journal of Epidemiology* (2021), pp. 1–15. ISSN: 0300-5771. DOI: 10.1093/ije/dyaa262.

[17] Stephen Burgess et al. "A robust and efficient method for Mendelian randomization with hundreds of genetic variants". In: *Nature Communications* 11.1 (2020). ISSN: 20411723. DOI: 10.1038/s41467-019-14156-4. URL: http://dx.doi.org/10.1038/s41467-019-14156-4.

[18] Stephen Burgess et al. "Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization". In: (2016), pp. 1–45. arXiv: 1606.03729. URL: http://arxiv.org/abs/1606.03729.

[19] Michael R. Shurin and Yuri S. Smolkin. "Immune-mediated diseases: Where do we stand?" In: *Advances in Experimental Medicine and Biology*. Vol. 601. 2007, pp. 3–12. ISBN: 9780387720043. DOI: 10.1007/978-0-387-72005-0_1.

[20] Nasa Sinnott-Armstrong et al. "Genetics of 35 blood and urine biomarkers in the UK Biobank". In: *Nature Genetics* (2021). ISSN: 1061-4036. DOI: 10.1038/s41588-020-00757-z. URL: http://dx.doi.org/10.1038/s41588-020-00757-z.

[21] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (Oct. 2018), pp. 203–209. ISSN: 14764687. DOI: 10.1038/s41586-018-0579-z. URL: https://doi.org/10.1038/s41586-018-0579-z.

[22] M ELLAURIE. "Platelet abnormalities in asthma and allergy*1". In: *Journal of Allergy and Clinical Immunology* 113.2 (Feb. 2004), S161. ISSN: 00916749. DOI: 10.1016/j.jaci.2004.01.009.

[23] Paul Stoll and Marek Lommatzsch. "Platelets in Asthma: Does Size Matter?" In: *Respiration* 88.1 (2014), pp. 22–23. ISSN: 0025-7931. DOI: 10.1159/000362798. URL: https://www.karger.com/Article/FullText/362798.

[24] ManalR Hafez et al. "Assessment of bronchial asthma exacerbation: the utility of platelet indices". In: *Egyptian Journal of Bronchology* 13.5 (Jan. 2019), p. 623. ISSN: 1687-8426. DOI: 10.4103/ejb.ejb_69_19. URL: http://www.ejbronchology.eg.net/text.asp?2019/13/5/623/276301.

[25] John W. Semple, Joseph E. Italiano, and John Freedman. *Platelets and the immune continuum*. Apr. 2011. DOI: 10.1038/nri2956. URL: https://pubmed.ncbi.nlm.nih.gov/21436837/.

[26] Fauzia Imtiaz et al. "Neutrophil lymphocyte ratio as a measure of systemic inflammation in prevalent chronic diseases in Asian population". In: *International Archives of Medicine* 5.1 (2012). ISSN: 17557682. DOI: 10.1186/1755-7682-5-2. URL: https://pubmed.ncbi.nlm.nih.gov/22281066/.

[27] Özlem Taşoğlu et al. "Is blood neutrophil-lymphocyte ratio an independent predictor of knee osteoarthritis severity?" In: *Clinical Rheumatology* 35.6 (June 2016), pp. 1579–1583. ISSN: 14349949. DOI: 10.1007/s10067-016-3170-8. URL: https://pubmed.ncbi.nlm.nih.gov/26780447/.

[28] Adil Can Gungen and Yusuf Aydemir. *The correlation between asthma disease and neutrophil to lymphocyte ratio.* Tech. rep. 0. 2017. URL: http://www.alliedacademies.org/research-journal-of-allergy-and-immunology/.

[29] Sophie Demarche et al. "Detailed analysis of sputum and systemic inflammation in asthma phenotypes: Are paucigranulocytic asthmatics really non-inflammatory?" In: *BMC Pulmonary Medicine* 16.1 (Apr. 2016), pp. 1–13. ISSN: 14712466. DOI: 10.1186/s12890-016-0208-2. URL: https://link.springer.com/articles/10.1186/s12890-016-0208-2%20https://link.springer.com/article/10.1186/s12890-016-0208-2.

[30] Ruurt A. Jukema, Tarek A.N. Ahmed, and Jean Claude Tardif. "Does low-density lipoprotein cholesterol induce inflammation? if so, does it matter? Current insights and future perspectives for novel therapies". In: *BMC Medicine* 17.1 (Nov. 2019), p. 197. ISSN: 17417015. DOI: 10.1186/s12916-019-1433-3. URL: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1433-3.

[31] Alexandros Tsoupras, Ronan Lordan, and Ioannis Zabetakis. *Inflammation, not cholesterol, is a cause of chronic disease.* May 2018. DOI: 10.3390/nu10050604. URL: /pmc/articles/PMC5986484/%20/pmc/articles/PMC5986484/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986484/.

[32] M. B. Fessler et al. "Relationship of serum cholesterol levels to atopy in the US population". In: *Allergy: European Journal of Allergy and Clinical Immunology* 65.7 (2010), pp. 859–864. ISSN: 13989995. DOI: 10.1111/j.1398-9995.2009.02287.x. URL: /pmc/articles/PMC4045486/%20/pmc/articles/PMC4045486/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4045486/.

[33] WILLIAM J. CALHOUN, JULIE SEDGWICK, and WILLIAM W. BUSSE. "The Role of Eosinophils in the Pathophysiology of Asthma". In: *Annals of the New York Academy of Sciences* 629.1 Advances in t (July 1991), pp. 62–72. ISSN: 0077-8923. DOI: 10.1111/j.1749-6632.1991.tb37961.x. URL: http://doi.wiley.com/10.1111/j.1749-6632.1991.tb37961.x.

[34] Dirk Smith et al. "A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma". In: *PLOS Genetics* 13.3 (Mar. 2017). Ed. by Tuuli Lappalainen, e1006659. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006659. URL: https://dx.plos.org/10.1371/journal.pgen.1006659.

[35] Maria Cappello et al. *Liver function test abnormalities in patients with inflammatory bowel diseases: A hospital-based survey.* June 2014. DOI: 10.4137/CGast.S13125. URL: /pmc/articles/PMC4069044/%20/pmc/articles/PMC4069044/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069044/.

[36] Jessika Barendregt et al. "Liver test abnormalities predict complicated disease behaviour in patients with newly diagnosed Crohn's disease". In: *International Journal of Colorectal Disease* 32.4 (Apr. 2017), pp. 459–467. ISSN: 14321262. DOI: 10.1007/s00384-016-2706-3. URL: https://link.springer.com/article/10.1007/s00384-016-2706-3.

[37] Emma H. Baker and Derek Bell. *Blood glucose: Of emerging importance in COPD exacerbations.* Oct. 2009. DOI: 10.1136/thx.2009.118638. URL: http://thorax.bmj.com/.

[38] Michael J. Berridge. *Vitamin D deficiency and diabetes.* Apr. 2017. DOI: 10.1042/BCJ20170042. URL: https://pubmed.ncbi.nlm.nih.gov/28341729/.

[39] P. Rocha-Pereira et al. "Erythrocyte damage in mild and severe psoriasis". In: *British Journal of Dermatology* 150.2 (Feb. 2004), pp. 232–244. ISSN: 00070963. DOI: 10.1111/j.1365-2133.2004.05801.x. URL: https://pubmed.ncbi.nlm.nih.gov/14996093/.

[40] Susana Coimbra et al. “The roles of cells and cytokines in the pathogenesis of psoriasis”. In: *International Journal of Dermatology* 51.4 (Apr. 2012), pp. 389–398. ISSN: 00119059. DOI: 10.1111/j.1365-4632.2011.05154.x. URL: http://doi.wiley.com/10.1111/j.1365-4632.2011.05154.x.

[41] Sibel Doğan and Nilgün Atakan. “Red blood cell distribution width is a reliable marker of inflammation in plaque psoriasis”. In: *Acta Dermatovenerologica Croatica* 25.1 (2017), pp. 26–31. ISSN: 18476538.

[42] C. Vahlquist, G. Michaelsson, and B. Vessby. “Serum lipoproteins in middle-aged men with psoriasis”. In: *Acta Dermato-Venereologica* 67.1 (1987), pp. 12–15. ISSN: 00015555.

[43] Mehdi Taheri Sarvtin et al. “Study of serum lipids and lipoproteins in patients with psoriasis”. In: *Journal of Mazandaran University of Medical Sciences* 23.98 (2013), pp. 173–177. ISSN: 17359260.

[44] Michael Holzer et al. “Psoriasis alters HDL composition and cholesterol efflux capacity”. In: *Journal of Lipid Research* 53.8 (Aug. 2012), pp. 1618–1624. ISSN: 00222275. DOI: 10.1194/jlr.M027367. URL: http://www.jlr.org.

[45] V. Sathiyapriya et al. “Evidence for the role of lipid peroxides on glycation of hemoglobin and plasma proteins in non-diabetic asthma patients”. In: *Clinica Chimica Acta* 366.1-2 (Apr. 2006), pp. 299–303. ISSN: 00098981. DOI: 10.1016/j.cca.2005.11.001.

[46] Miyoshi Fujita et al. “C-reactive protein levels in the serum of asthmatic patients”. In: *Annals of Allergy, Asthma and Immunology* 99.1 (July 2007), pp. 48–53. ISSN: 10811206. DOI: 10.1016/S1081-1206(10)60620-5.

[47] Natalie M. Niessen et al. “Neutrophilic asthma features increased airway classical monocytes”. In: *Clinical & Experimental Allergy* 51.2 (Feb. 2021), pp. 305–317. ISSN: 0954-7894. DOI: 10.1111/cea.13811. URL: https://onlinelibrary.wiley.com/doi/10.1111/cea.13811.

[48] Ibon Eguíluz-Gracia et al. “Monocytes accumulate in the airways of children with fatal asthma”. In: *Clinical and Experimental Allergy* 48.12 (Dec. 2018), pp. 1631–1639. ISSN: 13652222. DOI: 10.1111/cea.13265. URL: https://pubmed.ncbi.nlm.nih.gov/30184280/.

[49] A. Barry Kay. *The role of T lymphocytes in asthma.* 2006. DOI: 10.1159/000090230. URL: https://pubmed.ncbi.nlm.nih.gov/16354949/.

[50] Telugu A. Narasaraju et al. “Expression profile of IGF system during lung injury and recovery in rats exposed to hyperoxia: A possible role of IGF-1 in alveolar epithelial cell proliferation and differentiation”. In: *Journal of Cellular Biochemistry* 97.5 (Apr. 2006), pp. 984–998. ISSN: 0730-2312. DOI: 10.1002/jcb.20653. URL: http://doi.wiley.com/10.1002/jcb.20653.

[51] Yueh Ying Han et al. “Serum insulin-like growth factor-1, asthma, and lung function among British adults”. In: *Annals of Allergy, Asthma and Immunology* 126.3 (Mar. 2021), 284–291.e2. ISSN: 15344436. DOI: 10.1016/j.anai.2020.12.005.

[52] Mojtaba Shafiee et al. “Depression and anxiety symptoms are associated with white blood cell count and red cell distribution width: A sex-stratified analysis in a population-based study”. In: *Psychoneuroendocrinology* 84 (Oct. 2017), pp. 101–108. ISSN: 18733360. DOI: 10.1016/j.psyneuen.2017.06.021. URL: https://pubmed.ncbi.nlm.nih.gov/28697416/.

[53] Yuji Shimizu et al. “Short stature is an inflammatory disadvantage among middle-aged Japanese men”. In: *Environmental Health and Preventive Medicine* 21.5 (Sept. 2016), pp. 361–367. ISSN: 13474715. DOI: 10.1007/s12199-016-0538-y. URL: /pmc/articles/PMC5305990/%20/pmc/articles/PMC5305990/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5305990/.

[54] P. S. Sanmuganathan et al. “Aspirin for primary prevention of coronary heart disease: Safety and absolute benefit related to coronary risk derived from meta-analysis of randomised trials”. In: *Heart* 85.3 (Mar. 2001), pp. 265–271. ISSN: 13556037. DOI: 10.1136/heart.85.3.265. URL: www.heartjnl.com.

[55] Thomas M. MacDonald and Li Wei. "Is there an Interaction between the Cardiovascular Protective Effects of Low-Dose Aspirin and Ibuprofen?" In: *Basic ¡html_ent glyph="@amp;" ascii="&"/¿ Clinical Pharmacology ¡html_ent glyph="@amp;" ascii="&"/¿ Toxicology* 98.3 (Mar. 2006), pp. 275–280. ISSN: 1742-7835. DOI: 10.1111/j.1742-7843.2006.pto_371.x. URL: http://doi.wiley.com/10.1111/j.1742-7843.2006.pto%7B%5C_%7D371.x.

[56] Yanyan Zhu, Yuqing Zhang, and Hyon K. Choi. "The serum urate-lowering impact of weight loss among men with a high cardiovascular risk profile: The Multiple Risk Factor Intervention Trial". In: *Rheumatology* 49.12 (Dec. 2010), pp. 2391–2399. ISSN: 14620324. DOI: 10.1093/rheumatology/keq256. URL: https://pubmed.ncbi.nlm.nih.gov/20805117/.

[57] Marjolein Visser et al. "Elevated C-reactive protein levels in overweight and obese adults". In: *Journal of the American Medical Association* 282.22 (Dec. 1999), pp. 2131–2135. ISSN: 00987484. DOI: 10.1001/jama.282.22.2131. URL: https://pubmed.ncbi.nlm.nih.gov/10591334/.

[58] Rana H. Mosli and Hala H. Mosli. "Obesity and morbid obesity associated with higher odds of hypoalbuminemia in adults without liver disease or renal failure". In: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* 10 (Nov. 2017), pp. 467–472. ISSN: 11787007. DOI: 10.2147/DMSO.S149832. URL: /pmc/articles/PMC5687480/%20/pmc/articles/PMC5687480/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5687480/.

[59] Bogdan Pasaniuc and Alkes L. Price. *Dissecting the genetics of complex traits using summary association statistics*. Feb. 2017. DOI: 10.1038/nrg.2016.142. URL: /pmc/articles/PMC5449190/%20/pmc/articles/PMC5449190/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5449190/.

[60] Stephen Burgess, Adam S. Butterworth, and John R. Thompson. "Beyond Mendelian randomization: How to interpret evidence of shared genetic predictors". In: *Journal of Clinical Epidemiology* 69 (Jan. 2016), pp. 208–216. ISSN: 18785921. DOI: 10.1016/j.jclinepi.2015.08.001. URL: /pmc/articles/PMC4687951/%20/pmc/articles/PMC4687951/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687951/.

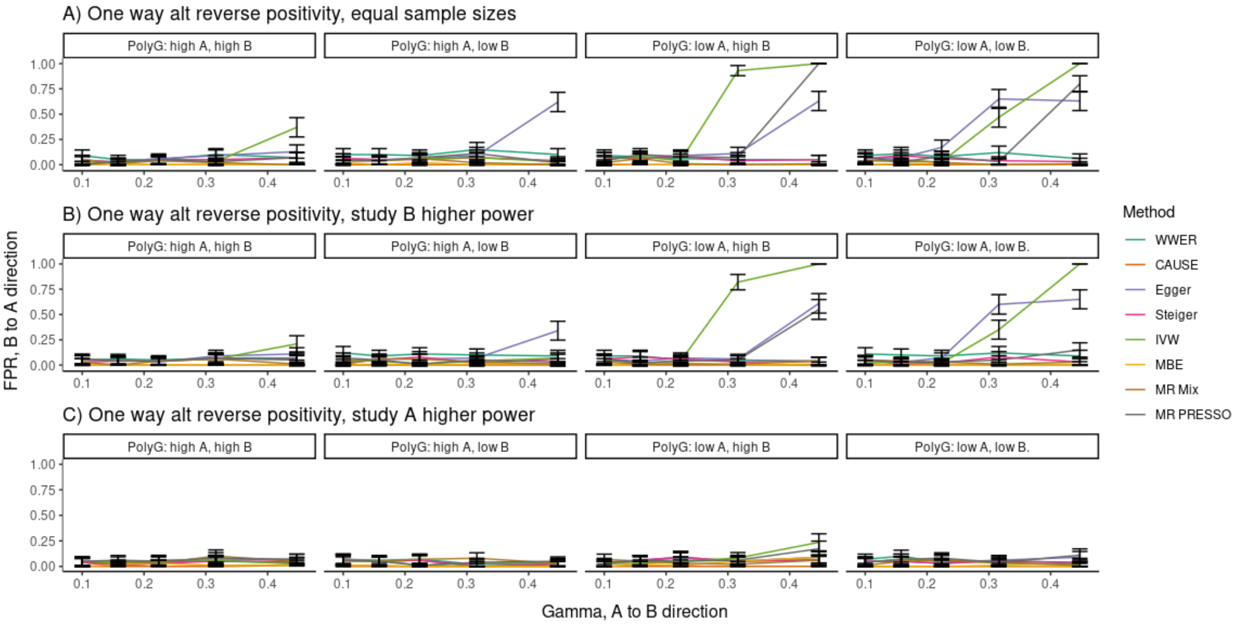


Figure S1: False positive rate in the reverse direction (B to A) when there is a causal effect in the A to B direction (x-axis). Settings correspond to the settings of Figure 3A-C. IVW, MR Presso and Egger often produce a false positive in the B to A direction when there is a strong effect in the A to B direction.

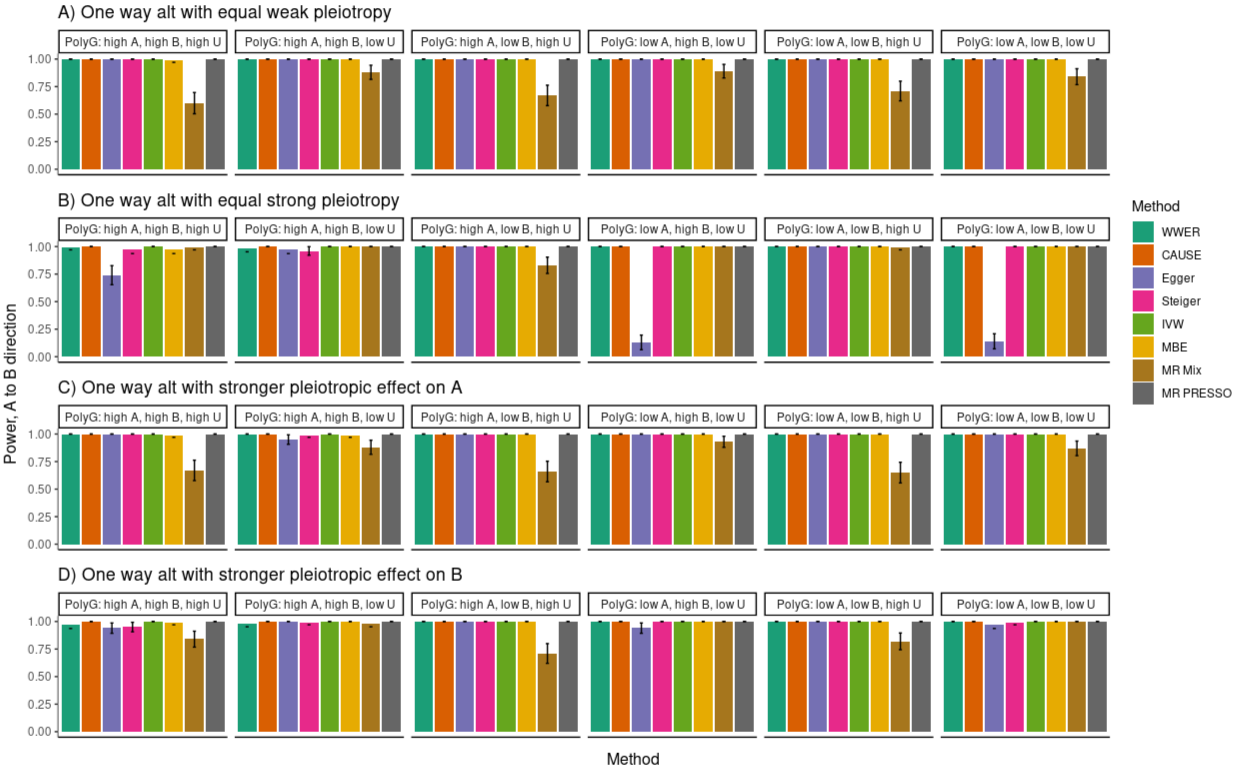


Figure S2: Power of each method for various settings when there is both a causal effect and correlated pleiotropy.

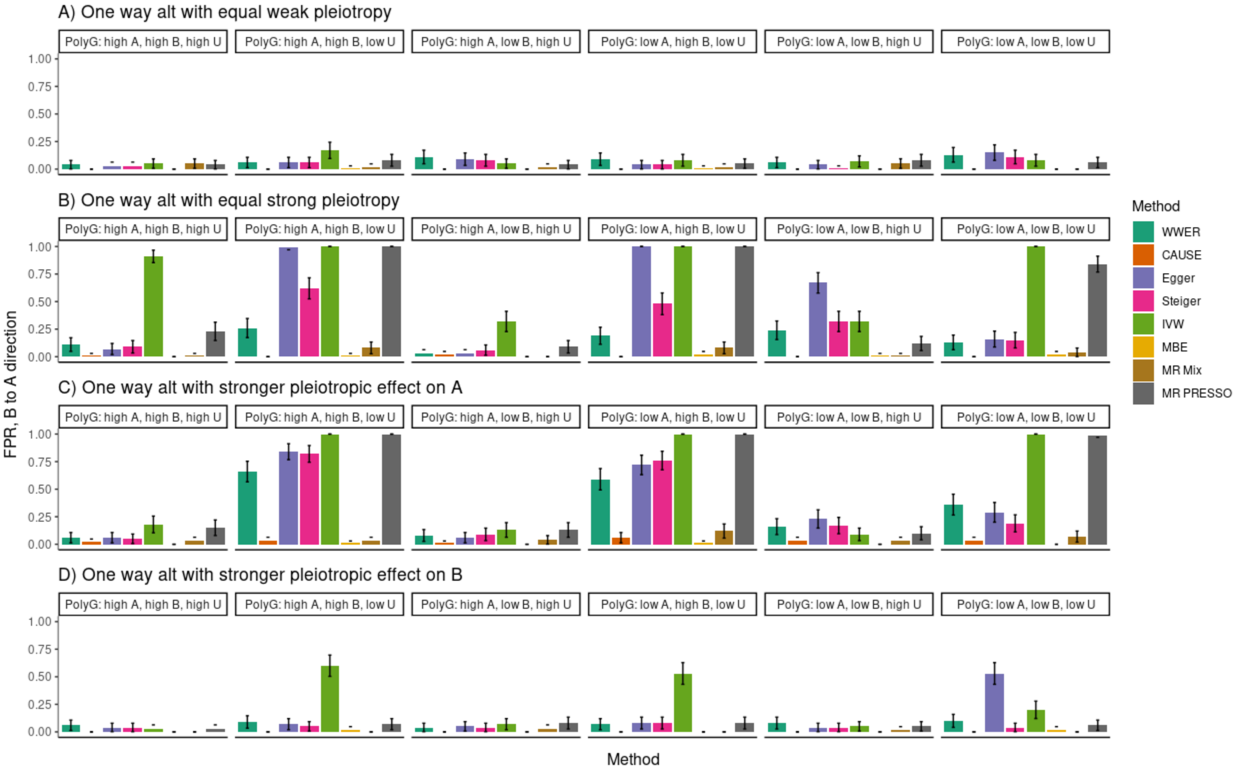
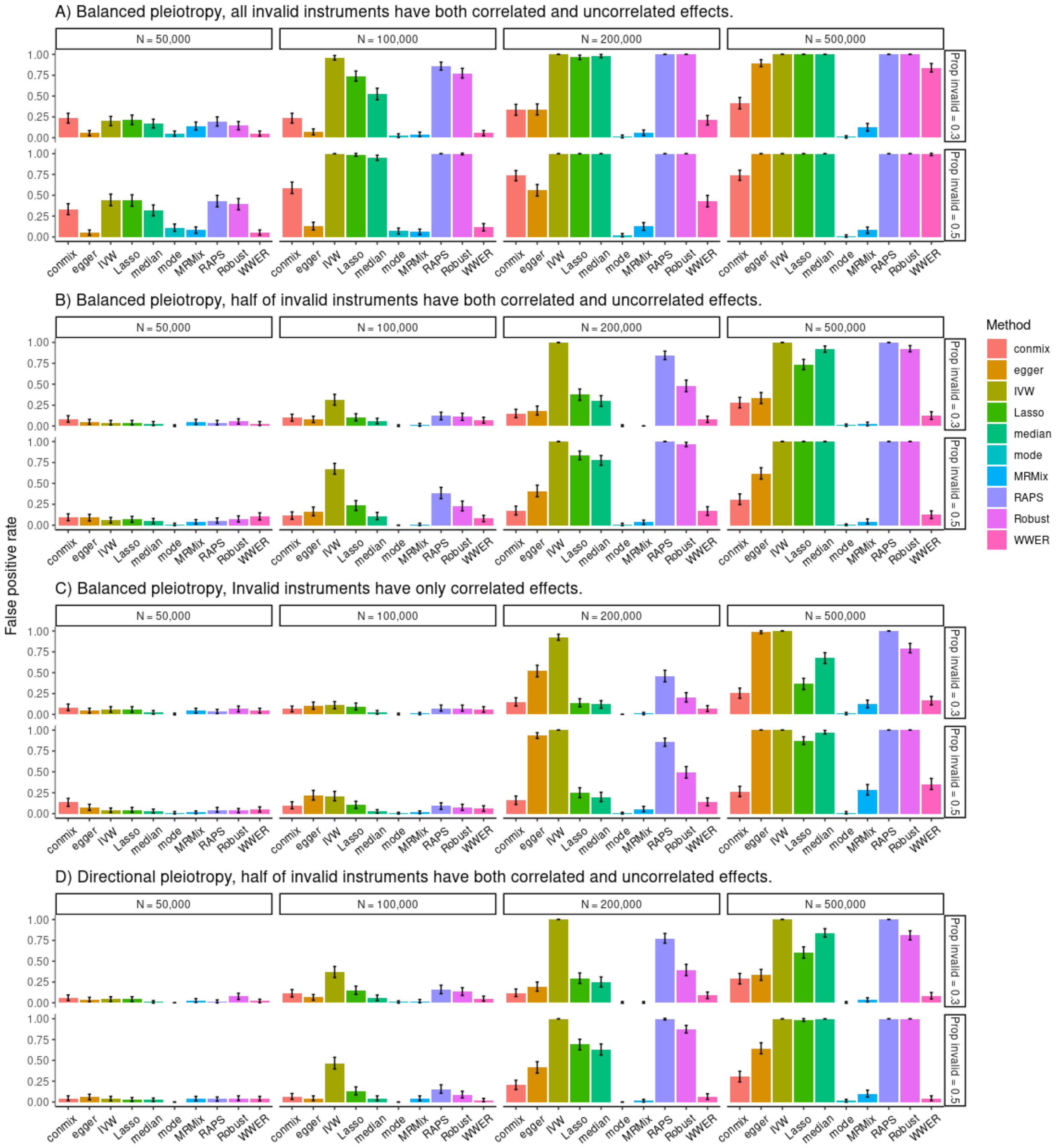


Figure S3: False positive rate in the reverse direction (B to A) when there is a causal effect in the forward direction (A to B) and correlated pleiotropy. Settings correspond to the settings in Figure S2. Both WWER and Stieger filtering show a substantial improvement over Egger regression and other traditional methods. Here WWER clearly outperforms Steiger filtering in several settings, although there are still two where it produces a high FPR.



A) Balanced pleiotropy, all invalid instruments have both correlated and uncorrelated effects.

B) Balanced pleiotropy, half of invalid instruments have both correlated and uncorrelated effects.

C) Balanced pleiotropy, Invalid instruments have only correlated effects.

D) Directional pleiotropy, half of invalid instruments have both correlated and uncorrelated effects.

Figure S4: False positive rate under the null using the model of [16].

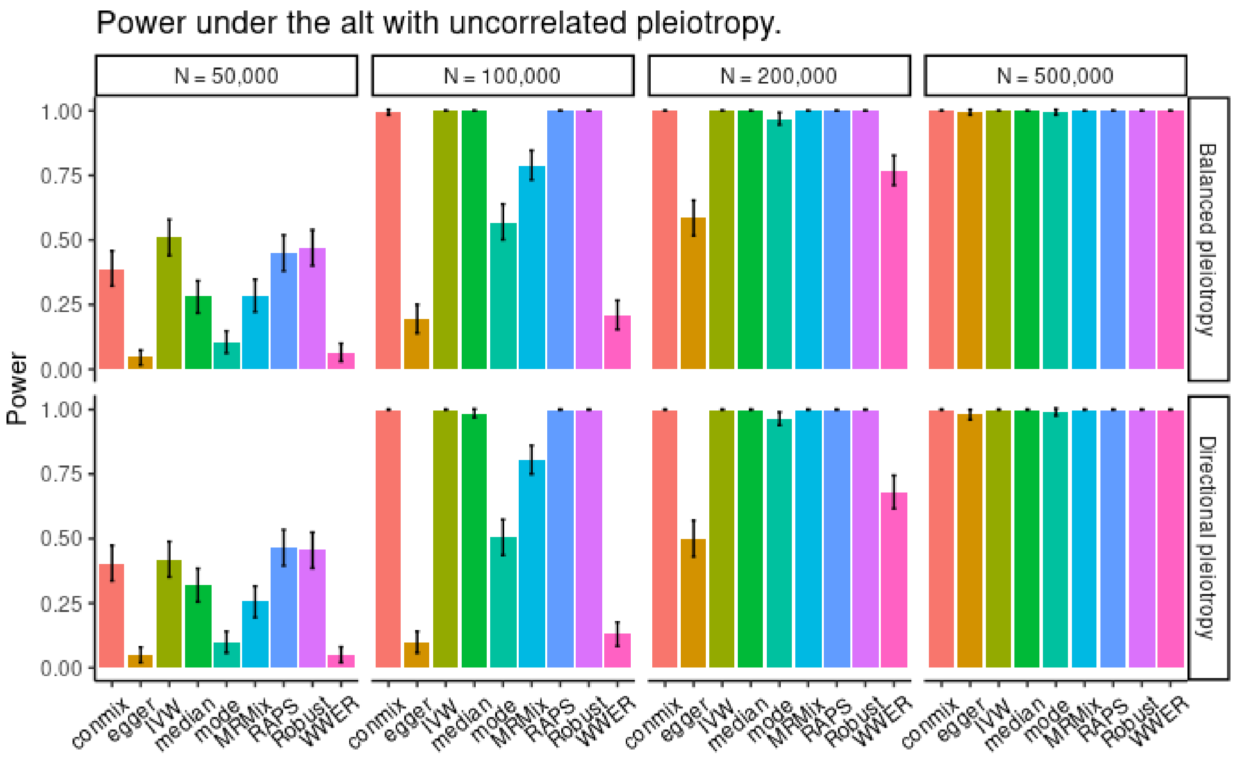


Figure S5: Power under the alt using the model of [16].

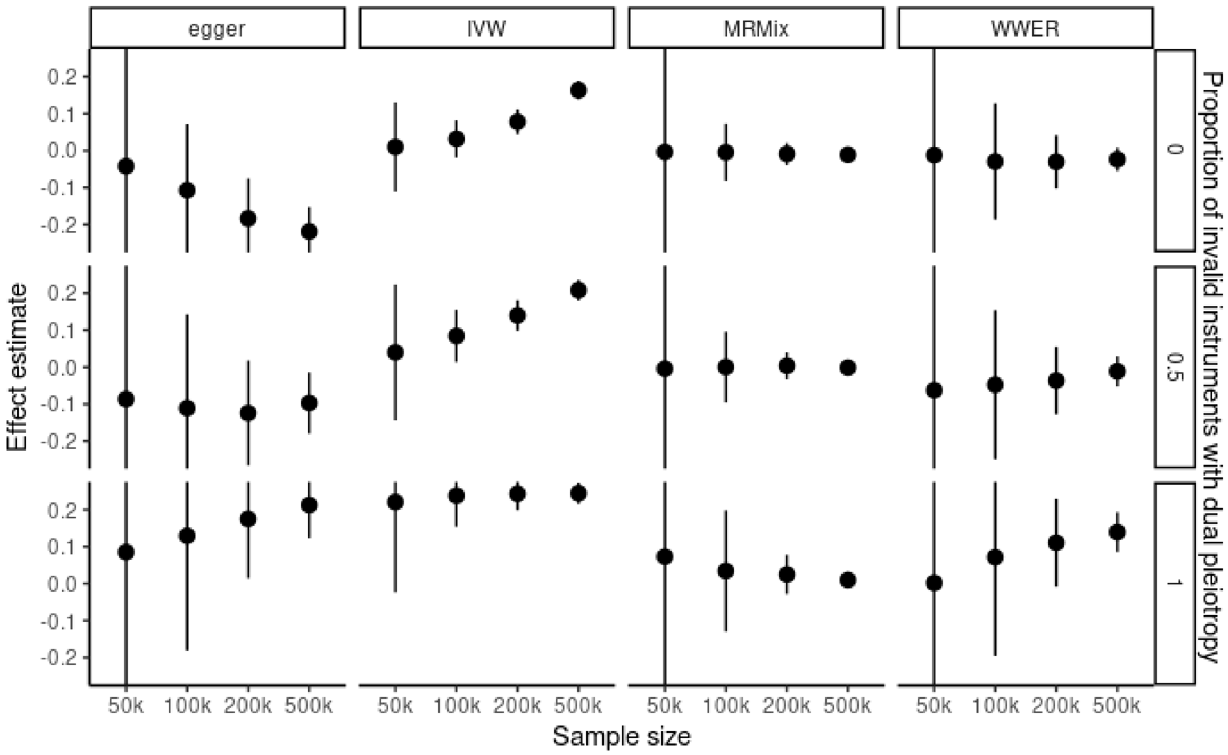


Figure S6: An experiment illustrating the change in effect estimate with increasing sample size.

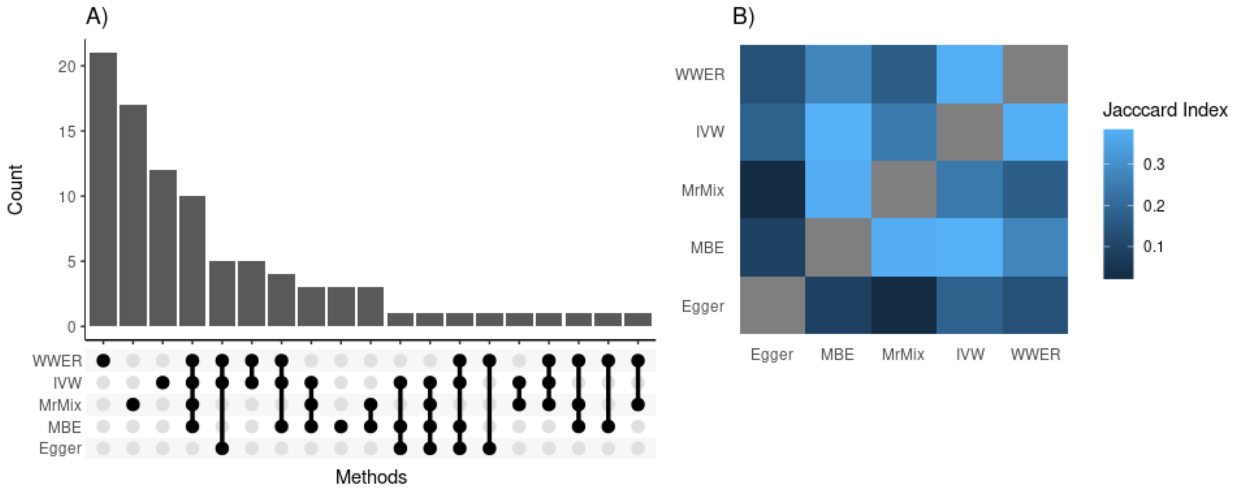


Figure S7: A comparison of WWER against other methods on real data. A) Upset plot comparing discoveries in the blood biomarker and IMID study across five MR methods. B) Jaccard coefficient matrix describing the similarity of the discovery sets between methods.

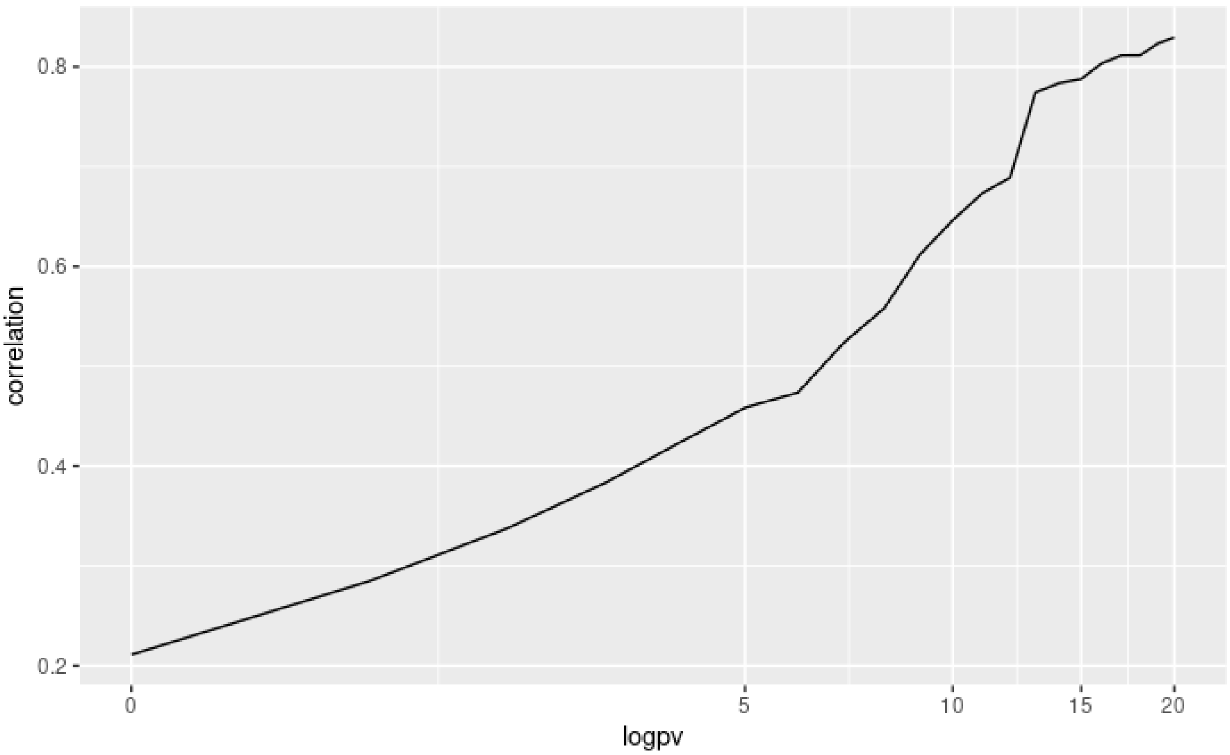


Figure S8: Correlation of the estimated causal effect with the genetic correlation, as a function of the significance of the estimated causal effect. The global correlation is low, but when the estimate of the causal effect is more significant it is also more similar to the estimated genetic correlation.

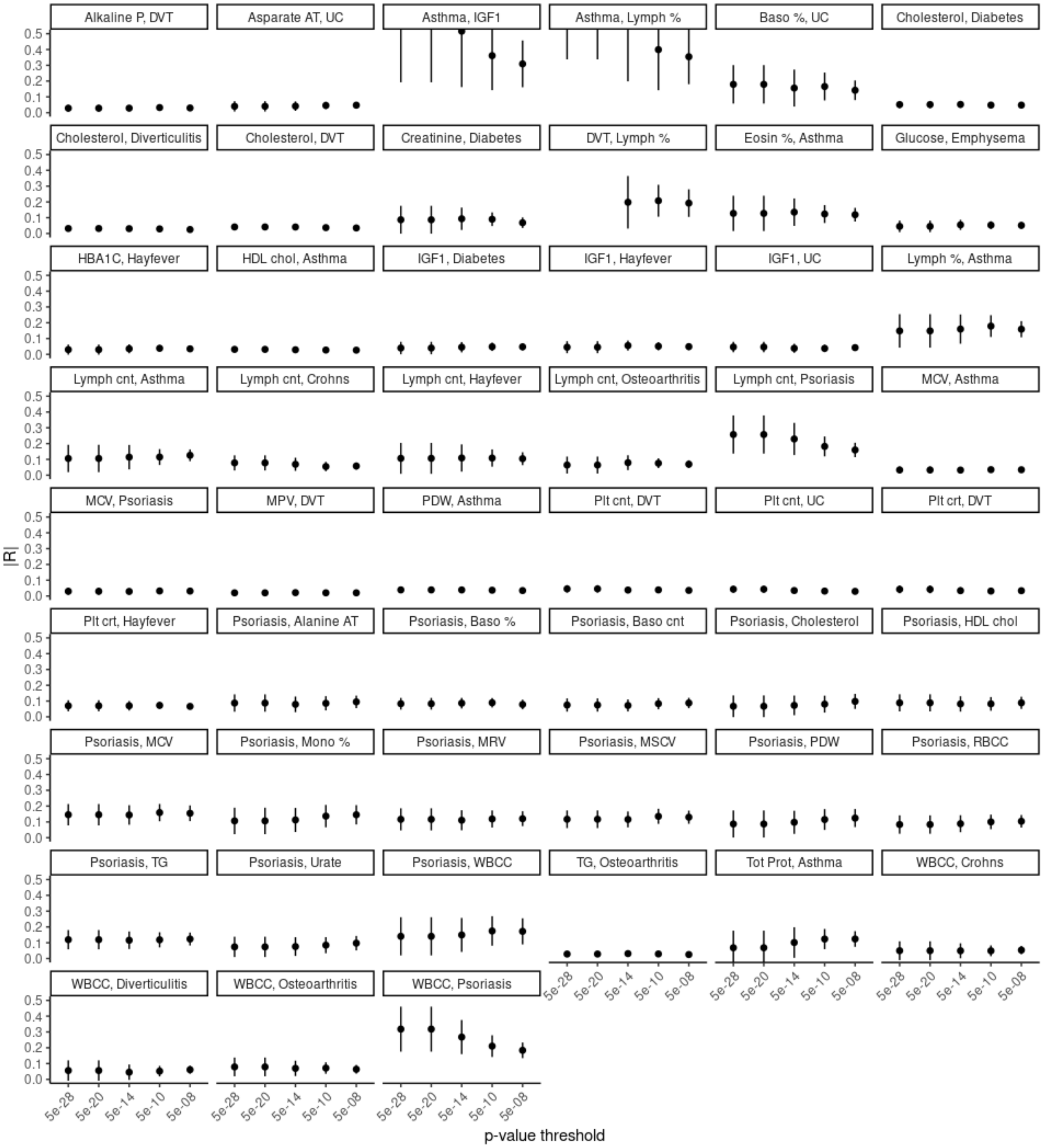


Figure S9: Effect estimate and standard error as a function of the SNP-inclusion threshold for FWER-significant pairs of phenotypes in out comparison against other methods. Pairs with substantially increasing effect estimates as a function of p-value threshold are considered false positives. These include the observed effect of total protein on asthma (row 8, column 5) and potentially of psoriasis on cholesterol (row 6, column 5).

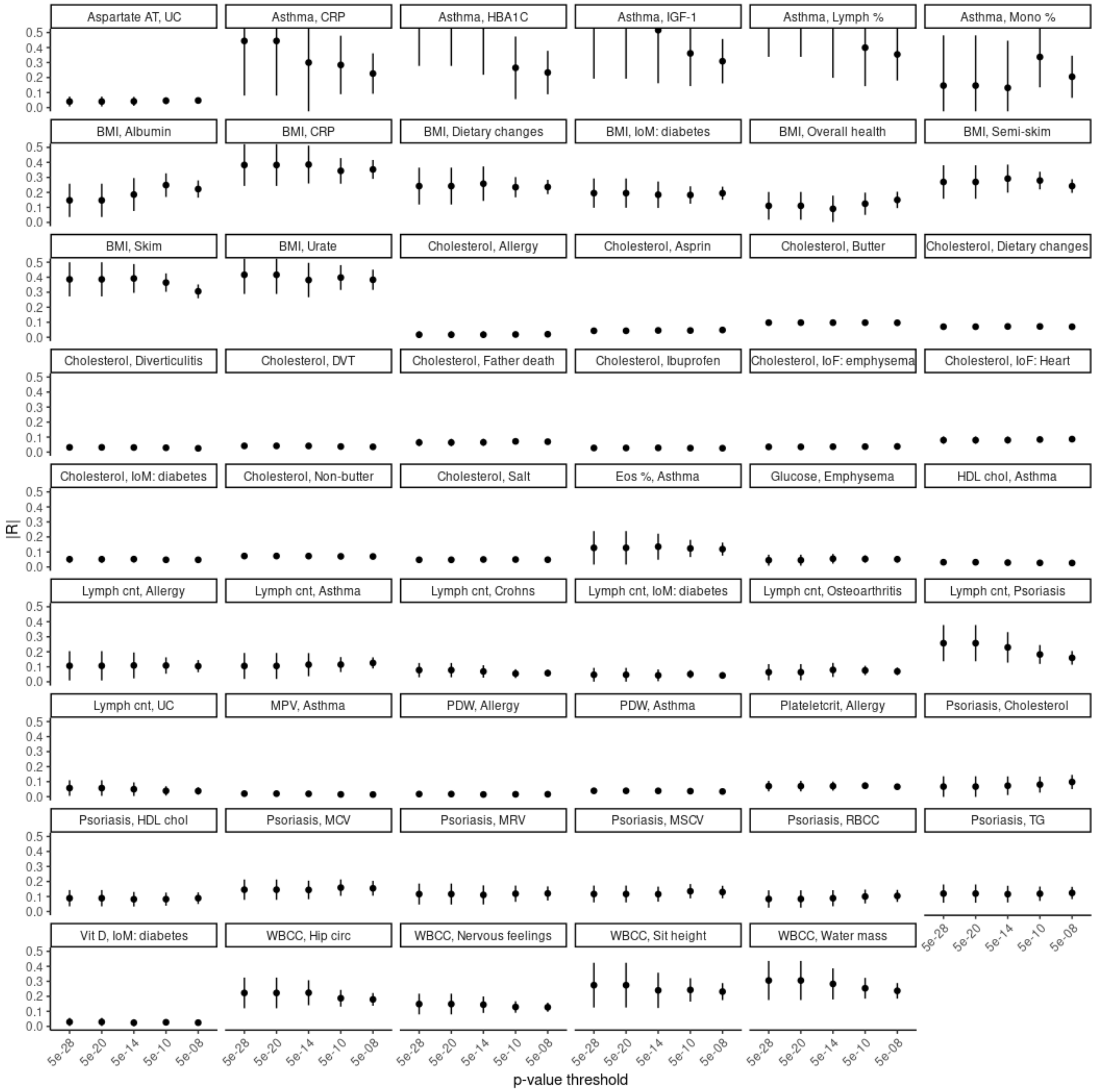


Figure S10: Effect estimate and standard error as a function of the SNP-inclusion threshold for all pairs mentioned in the main text. Pairs with substantially increasing effect estimates as a function of p-value threshold may be considered false positives. These include the observed effects of asthma on monocyte percentage (row 1, column 6), psoriasis on cholesterol (row 7, column 6), BMI on albumin (row 2, column 1) and possibly BMI on overall health rating (row 2, column 5).

Method	Description	Key Assumptions
Robust	IVW with robust regression	Invalid instruments are outliers
LASSO	IVW regression accounting for SNP-specific pleiotropy using an L_1 penalty	Invalid instruments are outliers
Robust	IVW with robust regression	Invalid instruments are outliers
Contamination mixture	Mixture model to classify variants as invalid or valid, fit with a profile likelihood	Balanced pleiotropy, ZEMPA, ratio estimates come from normal mixture distribution

Table S1: An overview of the methods considered in our comparisons.

Symbol	Definition	Class	Possible values	Example
M	Number of SNPs	Fixed	$[1, \infty)$	500,000
N_A	Sample size of study for phenotype A	Fixed	$[1, \infty)$	50,000
N_B	Sample size of study for phenotype B	Fixed	$[1, \infty)$	50,000
ν	Per-variance effect of U on A	Fixed	$[-1, 1]$	$\sqrt{0.2}$
η	Per-variance effect of U on B	Fixed	$[-1, 1]$	$\sqrt{0.2}$
γ	Per-variance effect of A on B	Fixed	$[-1, 1]$	$\sqrt{0.2}$
δ	Per-variance effect of B on A	Fixed	$[-1, 1]$	$\sqrt{0.2}$
q	Proportion of variants effecting U	Fixed	$[0, 1]$	0.3
r	Proportion of variants effecting A	Fixed	$[0, 1]$	0.3
s	Proportion of variants effecting B	Fixed	$[0, 1]$	0.3
h_A^2	Heritability of phenotype A	Fixed	$[0, 1]$	0.2
h_B^2	Heritability of phenotype B	Fixed	$[0, 1]$	0.2
h_U^2	Heritability of phenotype U	Fixed	$[0, 1]$	0.2
σ_A	Variance of the effect size distribution for A	Fixed	NA	1e−4
σ_B	Variance of the effect size distribution for B	Fixed	NA	1e−4
σ_U	Variance of the effect size distribution for U	Fixed	NA	1e−4
β_A	Effect of SNP on phenotype A, $\sim \mathcal{N}(0, \sigma_A)$	Random	$[-1, 1]$	$\sqrt{0.01}$
β_B	Effect of SNP on phenotype B, $\sim \mathcal{N}(0, \sigma_A)$	Random	$[-1, 1]$	$\sqrt{0.01}$
β_U	Effect of SNP on phenotype U, $\sim \mathcal{N}(0, \sigma_A)$	Random	$[-1, 1]$	$\sqrt{0.01}$
Z_A	Indicator that SNP effects A, sampled from $Bern(r)$	Random	$\{0, 1\}$	0
Z_B	Indicator that SNP effects B, sampled from $Bern(s)$	Random	$\{0, 1\}$	0
Z_U	Indicator that SNP effects U, sampled from $Bern(q)$	Random	$\{0, 1\}$	0

Table S2: A list of all parameters in our model, with examples and feasible ranges.

Supplemental Methods

Simulating from the proposed model

Our simulation framework (Figure 1) has 19 free parameters and 3 parameters that are a deterministic function of the others, see Table 9 for an overview. We let N_A and N_B represent the sample sizes of the studies for phenotypes A and B, respectively, and M be the number of SNPs. The effects of each phenotype on each other are given by ν, η, γ, and δ for the effect of U on A, U on B, A on B and B on A, respectively. q, r, and s control the proportion of variants effecting phenotypes U, A, and B, respectively, and h_A^2, h_B^2 and h_U^2. As we describe below, these parameters allow us to determine the variance of the effect size distribution for each phenotype, which we represent by σ_A^2, σ_B^2 and σ_U^2. We sample the M-vector of effect sizes $\beta_\bullet \sim \mathcal{N}(0, \sigma_\bullet^2)$ and determine which traits each SNP affects by sampling M-vector indicator variables $Z_\bullet \sim \text{Bern}(\pi_\bullet)$, with $\bullet$ representing one of A, B or U, and $\pi_\bullet$ being the proportion of SNPs with non-zero direct effects on $\bullet$.

We derive the variance of the phenotypes in this model. The phenotype values are given by

$$U = \boldsymbol{X}\beta_U \circ Z_U + \epsilon_U, \tag{6}$$

$$A = U\nu + B\delta + \boldsymbol{X}\beta_A \circ Z_A + \epsilon_A, \tag{7}$$

$$B = U\eta + A\gamma + \boldsymbol{X}\beta_B \circ Z_B + \epsilon_B, \tag{8}$$

where Z's represent indicator variables that the SNP affects that trait, sampled as indicated above, $\circ$ indicates vector element-wise (Hadamard) multiplication, and bolding represents matrices. Here $\epsilon_\bullet \sim \mathcal{N}(0, e_\bullet) = \mathcal{N}(0, 1 - h_\bullet^2)$ is the residual (environmental) contribution to each phenotype. Plugging the expressions for U and B into A and solving for A we find that

$$A = \frac{\nu + \eta\delta}{1 - \gamma\delta}(X\beta_U Z_U + \epsilon_U) + \frac{\delta}{1 - \gamma\delta}(X\beta_B Z_B + \epsilon_B) + \frac{1}{1 - \gamma\delta}(X\beta_U Z_A + \epsilon_A), \tag{9}$$

$$= R_{UA}(X\beta_U Z_U + \epsilon_U) + R_{BA}(X\beta_B Z_B + \epsilon_B) + R(X\beta_U Z_A + \epsilon_A) \tag{10}$$

where we have introduced the shorthand $R_{UA} = (\nu + \eta\delta)/(1 - \gamma\delta)$, $R_{BA} = \delta/(1 - \gamma\delta)$ and $R = R_{AA} = R_{BB} = 1/(1 - \gamma\delta)$ to represent the *total* causal effect of U on A, B on A and A on itself, respectively. It is important to notice that, because A can affect itself via a bi-directional effect on B that propagates back to A, $R = 1$ if and only if at least one of γ or δ is 0.

Let $g_\bullet = \text{Var}(X\beta_\bullet Z_\bullet)$ be the variance component of $\bullet$ contributed by direct genetic effects. By mirroring the above derivation for B, the variance of phenotypes A and B can be broken down as

$$\operatorname{Var} A = R_{UA}^2(g_U + e_U) + R_{BA}^2(g_B + e_B) + R^2(g_A + e_A), \tag{11}$$

$$\operatorname{Var} B = R_{UB}^2(g_U + e_U) + R_{AB}^2(g_A + e_A) + R^2(g_B + e_B) \tag{12}$$

Thus if the parameters are set such that $\operatorname{Var} A = \operatorname{Var} B = 1$, then the heritabilities are given by

$$h_A^2 = R_{UA}^2 g_U + R_{BA}^2 g_B + R^2 g_A, \tag{13}$$

$$h_B^2 = R_{UB}^2 g_U + R_{AB}^2 g_A + R^2 g_B \tag{14}$$

This is quite natural; the heritability of A is interpretable as the variance in A explained by U times the genetic component of U, plus the variance in A explained by B times the genetic component of B, plus the effect of A on itself times the genetic component of A (and likewise for B). The reverse is also true: if $h_\bullet^2$ is given as above and $e_\bullet = 1 - h_\bullet^2$, then the variances of each phenotype ($\operatorname{Var} \bullet$) are 1.

Of course, the variances of A and B will not be 1 for arbitrary settings of the parameters. Next, we show how to constrain them to be 1 so that parameters can be easily set on a per-variance scale. We manipulate (13) and (14) to determine g_A and g_B and thus σ_A and σ_B. Specifically, let

$$Q_A = h_A^2 - g_U R_{UA}^2 = g_B R_{BA}^2 + g_A R^2, \tag{15}$$

$$Q_B = h_B^2 - g_U R_{UB}^2 = g_A R_{AB}^2 + g_B R^2 \tag{16}$$

solving for g_A and g_B gives

$$g_A = \frac{R^2 Q_A - R_{BA}^2 Q_B}{R^4 - R_{AB}^2 R_{BA}^2}, \tag{17}$$

$$g_B = \frac{R^2 Q_B - R_{AB}^2 Q_A}{R^4 - R_{AB}^2 R_{BA}^2} \tag{18}$$

so that σ_A^2 and σ_B^2 can be determined

$$\sigma_A^2 = \frac{g_A}{M}, \tag{19}$$

$$\sigma_B^2 = \frac{g_B}{Ms} \tag{20}$$

Under the null, $\gamma = \delta = 0$, and therefore $R_{UA} = \nu$ and $R_{UB} = \eta$, while $R_{BA} = R_{AB} = 0$ and $R = 1$. In this case, the genetic covariance between A and B is $\text{Cov}(A, B) = h_U^2 \eta\nu$. Therefore the genetic correlation is

$$\rho_G = \frac{h_U^2 \eta\nu}{\sqrt{h_A^2 h_B^2}} = \frac{g_U^2 \eta\nu}{\sqrt{(g_U \nu^2 + g_A)(g_U \eta^2 + g_B)}} \tag{21}$$

Relationship to other models

Our model is very flexible and therefore contains other recently proposed models as special cases. Here we describe the relationship between our model and those used in LCV and CAUSE. Neither of these explicitly model the genetic architecture of the unobserved trait, preferring to tie it into the architecture of the observed traits. LCV is agnostic as to which trait is the exposure and which trait is the outcome, whereas CAUSE explicitly models one trait as the exposure (M, in their notation) and the other as the outcome (Y in their notation). For clarity when comparing to CAUSE we will use A as the exposure and B as the outcome, but it is important to remember our model is also agnostic to which trait is the exposure.

In the LCV model, under the null, the genetic correlation is $\rho_G = \eta\nu$, which we can arrive at by setting $h_U^2 = h_A^2 = h_B^2$. Their method attempts to quantify the deviation from a symmetric effect of the latent variable on the two observed variables, therefore the null case corresponds to $\nu = \eta$ ($q_1 = q_2$ in their notation). Finally, their settings focus on the case where the effect distribution of the SNPs acting on the observed traits is the same for SNPs acting directly and via the latent variable. In our model we can enact this assumption via the constraints $\sigma_u^2 \nu^2 = \sigma_a^2$ and $\sigma_u^2 \eta^2 = \sigma_b^2$. Finally, we assume that the SNPs effect a single trait in expectation, that is $q + r + s = 1$. Under the null, $h_A^2 = g_U \nu^2 + g_A$. Using the assumption that $h_U^2 = h_A^2$, we have that $g_U = g_U \nu^2 + g_A$. Rearranging and simplifying, we have $Mq\sigma_U^2(1 - \nu^2) = Mr\sigma_A^2 = Mr\sigma_U^2 \nu^2$ and thus $r = q(1 - \nu^2)/\nu^2$ (and likewise for s). Plugging into $q + r + s = 1$ and simplifying leads to

$$q = \frac{\nu^2}{2 - \nu^2} \tag{22}$$

The CAUSE model also assumes that $\sigma_u^2 \nu^2 = \sigma_a^2$. This is represented by the 1 on the arrow from the latent factor to the exposure in their model, indicating that SNPs have the same effect distribution on the exposure when acting via the latent variable or directly. In that model, the proportion of variants effecting the unobserved variable q controls the magnitude of the genetic correlation which then implicitly determines the heritibility of the hidden variable h_U^2. Again using $h_A^2 = g_U^2 \nu^2 + g_A^2$ and substituting $\sigma_u^2 \nu^2 = \sigma_a^2$ we find $h_A^2 = Mq\sigma_U^2 \nu^2 + Mr\sigma_U^2 \nu^2$. Solving for σ_U^2 and using the fact that $h_U^2 = Mq\sigma_U^2$ we find

$$h_U^2 = \frac{Mq}{Mq + Mr} \frac{h_a^2}{\nu^2} \tag{23}$$