Learning pleiotropic signatures to construct a causal map of the human phenome

29 July 2017

Gibran Hemani[1], Jack Bowden[1], Philip Haycock[1], Jie Zheng[1], Oliver Davis[1], Peter Flach, Tom Gaunt[1,*], George Davey Smith[1,*]

1. MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, School of Social and Community Medicine, Bristol, UK

Correspondence to: g.hemani@bristol.ac.uk

* Equal contribution

Abstract

A major application for genome-wide association studies (GWAS) has been the emerging field of causal inference using Mendelian randomization (MR), where the causal effect between a pair of traits can be estimated using only summary level data. MR depends on SNPs exhibiting vertical pleiotropy, where the SNP influences an outcome phenotype only through an exposure phenotype. Issues arise when this assumption is violated due to SNPs exhibiting horizontal pleiotropy, and many methods have been developed in an attempt to address this. Here we show that the mechanisms that underlie horizontal pleiotropy are numerous, and that instrument selection will be increasingly liable to selecting invalid instruments as GWAS sample sizes continue to grow. We have developed a mixture of experts machine learning approach (MR-MoE 1.0) that improves on both power and false discovery rates over unselective use of any existing methods. Using the approach, we systematically estimated the causal effects amongst 2407 phenotypes, generating a working draft of the causal map of the human phenome. Almost 90% of causal estimates indicated some level of horizontal pleiotropy. All results are organised into a publicly available graph database (http://eve.mrbase.org).

Introduction

Mendelian randomization (MR) (1,2) exploits genetic pleiotropy to infer the causal relationships between phenotypes. Suppose that one trait (the exposure) causally influences another (the outcome). If a SNP influences the outcome through the exposure then the SNP is exhibiting vertical pleiotropy. Such a genetic variant is considered to be an instrumental variable for the if it only influences the outcome through the exposure (the exclusion restriction). It can be exploited to mimic a randomised controlled trial, enabling a causal estimate to be made by comparing the outcome phenotypes between those individuals that have the exposure-increasing allele against those who do not. Multiple independent genetic variants for a particular exposure can be used jointly to improve causal inference, because a) each variant represents an independent natural experiment, and an overall causal estimate can be obtained by meta-analysing the single estimates from each instrument; and b) potential bias arising from the exclusion restriction principal can be detected or corrected by evaluating the consistency of effects across instruments (3–8).

Genome-wide association studies (GWAS) have identified genetic instrumental variables for thousands of phenotypes (9). Recent developments in MR have enabled knowledge of instrumental variables to be applied using only summary level data (known as two-sample MR, 2SMR) (10). Here, in order to infer the causal effect of an exposure on an outcome all that is required is an estimate of the genetic effects of the instrumenting SNP on the exposure, and the corresponding estimate of the effect on the outcome. This has three major advantages. First, GWAS summary data are non-disclosive and often publicly available. Second, causal inference can be made between phenotypes even if they have not been measured in the same samples, limiting possible causal estimates only by the availability of GWAS summary data for the traits in question (11).

Third, we can view the causal inference problem through the simple and widely understood prism of a meta-analysis.

Problems with obtaining unbiased causal effects can arise, however, if the genetic instruments exhibit horizontal pleiotropy (HP), where they influence the outcome through a pathway other than the exposure (2). The extent of this phenomenon is not to be understated, and many methods have been developed that attempt to reliably obtain unbiased causal estimates under specific models of HP (3–6,8,12). It is considered best practice to report estimates from all available methods as sensitivity analyses when presenting causal estimates because different pleiotropic models can then be evaluated. However it is also desirable to consolidate across methods under some circumstances. First, it could be of interest to make causal effect estimates for thousands of traits, in which case a critical evaluation of each causal estimate of interest may not be possible or convenient. Second, though the IVW approach is most statistically powerful under no HP, it can have high false negative or low true positive rates in the presence of HP compared to other methods. Pleiotropy has been hypothesised to be universal (13), though the degree to which this may be the case is contested (14,15), hence defaulting to the IVW method in the first instance and using other methods as sensitivity analyses may not be appropriate. Third, if different methods disagree it is not possible to know which is correct because the true nature of HP exhibited by the instruments is not known. Fourth, the available methods do not cover all possible models of HP, and therefore an automated method for instrument selection may be necessary.

In this paper we introduce two innovations towards improving the reliability of MR estimates. First, we introduce an approach to discard genetic variants that are likely to be invalid. Second, we hypothesised that characteristics of the summary data could indicate which method would be most reliable, and we introduce new machine learning approaches that attempt to automate both instrument and method selection. Using curated GWAS summary data for thousands of phenotypes (11), we use these new methods to construct a graph of millions of causal estimates. Motivated by the recent avalanche (16) of 2SMR publications, often with repeating (17,18) or contradicting (19,20) conclusions despite using the same data, we developed a graph database to represent these estimates in a consistent manner. We consider this to be a 'working draft' of the causal map of the human phenome, but raise caution throughout that its interpretation is far from straightforward, and that future corrections and refinements will inevitably follow as data grows and 2SMR methods evolve.

Methods

GWAS summary data and their use in 2SMR

The use of summary data in two-sample MR is described in detail elsewhere (11). A brief outline of the procedure is as follows. First, genetic instruments for the exposure trait need to be identified - those SNPs with $p < 10^{-8}$ are retained in order to ensure that the first assumption of MR (that the instrument associates with the exposure) is generally satisfied. We collect their effect sizes, standard errors and effect alleles for the association with the exposure trait. These can be obtained manually from complete summary data or from curated lists of significant GWAS associations. Next, the effects of those SNPs on the outcome need to be obtained, typically necessitating complete summary data because these SNPs will generally have not have reached genome-wide significance for other traits (and therefore be present in curated catalogues). At its most simple implementation, the regression of the SNP-exposure effect sizes against the SNP-outcome effect sizes, with greater weight afforded to those SNPs with smaller SNP-outcome standard errors, provides the estimate of the causal effect of the exposure on the outcome. This is known as the inverse-variance weighted (IVW) method.

Given summary data for a large number of traits, it is straightforward to exhaustively analyse the causal relationships of every trait against every other trait for which there are sufficient summary data available. Supplementary table 1 provides a list of all traits that have available GWAS summary data, indicating if they have complete summary data (in which case they can be used as both exposure traits and outcome traits), or if they only have significant associations (in which they can only be used as exposure traits).

MR methods and their assumptions

In this paper we consider three main classes of MR estimation. Full details for each approach have been described previously. A summary of the methods is given in Supplementary table 2.

Mean-based methods: Here we consider four nested models (5). The inverse variance weighted (IVW) fixed effects meta-analysis approach assumes that variants exhibit no HP. IVW random effects meta-analysis relaxes the HP assumption, allowing it to be present but balanced - such that it only leads to increased heterogeneity around the regression and not introducing bias. Fixed effects Egger regression (3) relaxes the HP assumption further by allowing a non-zero intercept which essentially allows horizontal pleiotropy to be directional, where it systematically influences the outcome in a specific direction. Random effects Egger regression allows heterogeneity around the directional HP (5), as long as the HP effects are not correlated with the SNP-exposure effects (also known as the INSIDE assumption)(3).

The Rucker framework (21), adapted to MR (5) uses estimates of heterogeneity in the IVW and Egger frameworks to navigate between these nested models. A jackknife approach (random selection with replacement of the complete set of instruments) can be used to obtain a sampling distribution for the model estimate amongst these four variations. Using 1000 rounds of jackknife estimates, we can obtain a final estimate using the mean or the median of the distribution. We only use the jackknife approach for associations where there are 15 or more variants in order to avoid saturating the possible number of instrument combinations.

The four nested models (IVW fixed effects, IVW random effects, Egger fixed effects, Egger random effects) plus the three Rucker estimates (point estimate, mean of the jackknife, median of the jackknife) provide seven mean based estimators.

Median-based methods: An alternative approach is to take the median effect of all available instruments (4,22). This has the advantage that up to half the instruments can be invalid, and the estimate will remain unbiased. Developing the approach further to allow stronger instruments to contribute more towards the estimate can be obtained by obtaining the median of the weights of each instrument. The penalised weighted median estimator introduces a further weight to the instruments, penalising any instrument that contributes substantially towards the heterogeneity statistic. Together, this provides three median based estimators. Other estimation strategies not considered here, such as LASSO regression, have also been developed for the situation where at least half of the instruments are valid (23).

Mode-based methods: Similar to the median, the mode based estimator clusters the instruments into groups based on similarity of causal effects, and returns the final causal effect estimate based on the cluster that has the largest number of instruments (6). There are four implementations of this method: the simple and the weighted mode, each weighted with or without the assumption of no measurement error in the exposure estimates (NOME). The simple mode is the unweighted mode of the empirical density function of causal estimates, whereas the weighted mode is weighted by the inverse variance of the outcome effect. The bandwidth parameter was set to 1 by default.

Instrument selection

Top hits

The simplest approach to selecting instruments for performing MR is to take SNPs that have been declared significant in the published GWAS for the exposure. This typically involves obtaining SNPs that surpass $p < 5 \times 10^{-8}$, using clumping to obtain independent SNPs, and then replicating in an independent sample. These results are often recorded in public GWAS catalogs. Alternatively the clumping procedure can be performed using complete summary data in MR-Base (11). We call this the “top hits” strategy.

Steiger filtering

With genome-wide association studies growing ever larger, the statistical power to detect significant associations that may be influencing the trait downstream of many other pathways increases. For example, if a SNP g_A

influences trait A, and trait A influences trait B, then a sufficiently powered GWAS will identify the g_A as being significant for trait B (Figure 1a). Using g_A as an instrument to test the causal effect of A on B is perfectly valid. But in the (incorrectly hypothesised) MR analysis of trait B on trait A could erroneously result in the apparent causal association of B on A. If g_A is only one of many known instruments for B, amongst which some are valid, it is to the advantage of the researcher to exclude g_A from the analysis.

An approach to inferring the causal direction between phenotypes recently developed (7) uses the following basic premise. If trait A causes trait B then

$$\sum_{i=1}^{M} cor(g_i, A)^2 > \sum_{i=1}^{M} cor(g_i, B)^2$$

because the $cor(g_i, B)^2 = cor(A, B)^2 cor(g_i, A)^2$. This simple inequality will not hold in some cases, for example $\rho_{x,x_o} < \rho_{x,y}\rho_{y,y_o}$ where ρ_{x,x_o} and ρ_{y,y_o} are the precision of the measurements of the x and y. Some combinations of confounding effects can also distort the $\rho_{g,x}$ and $\rho_{g,y}$ parameters, as has been discussed in detail previously. However, we use it here as a computationally inexpensive and approximate method to identify variants that are likely to be invalid. Steiger's Z-test of correlated correlations (24) can be used to formally test the extent to which the two correlations are statistically different.

Here we adapt this approach to automatically filter SNPs that are liable to be invalid (Figure 1a). Other methods have been developed to identify invalid instruments for the purposes of exclusion from MR analyses (8,12,25), based on the notion that outlying instruments are likely to be exhibiting horizontal pleiotropy. Conversely, we primarily developed Steiger filtering to identify those instruments that are likely to be arising due to reverse cause.

The Steiger test is applied to each variant in turn and we exclude any g_A for which $cor(g_i, A)^2 > cor(g_i, B)^2$, indicating that it is unlikely to primarily associate with B relative to A. Similarly, for SNPs that influence confounders of A and B or exhibit horizontal pleiotropy, the difference in $cor(g_i, A)^2$ and $cor(g_i, B)^2$ will be reduced, increasing the likelihood of the SNP being excluded because the Steiger Z-test is less likely to be significant. Hence we also exclude any g_A for which the Steiger Z-test has $p > 0.05$.

To estimate $cor(g, x)^2$, if x is continuous we obtain the F-statistic from the reported p-value and sample size and then $cor(g, x)^2 = \frac{F}{N-2-F}$. If x is binary then we estimate the variance of the underlying liability explained by the SNP, $cor(g, x)^2 = \frac{V_a}{V_a+V_e}$. Here, $V_e = \pi^2/3$, and $V_a = 2\beta^2 p(1-p)$, where β is the log odds ratio and p is the allele frequency of the SNP in the population (26). p can be estimated using the allele frequency of the SNP in an ascertained sample by deriving the 2×2 contingency table from the odds ratio e^{β}, allele frequency in the ascertained sample $p_{cc}, and number of cases N_1 and controls N_0. We exclude any SNPs that do not have the correct causal direction for the hypothesis, and do not have Steiger test $p < 0.05$.

Competitive mixture of experts

We consider the 14 MR methods described above, for which instruments can be supplied using two instrument selection strategies, leading to 28 methods in total (Supplementary table 2). In the context of this analysis, each method is considered to be an 'expert', taking a set of SNP-exposure and SNP-outcome effect sizes and their standard errors as inputs. Our objective is to select the expert most likely to be correct for a specific MR analysis. A overview of the approach is shown in Figure 2.

The mixture of experts method is a machine learning approach (27) which seeks to divide a parameter space into subdomains, such that a particular expert is used primarily for problems that reside in a subdomain most suited to that expert. In this case our objective is to identify characteristics of the SNP-exposure and SNP-outcome associations for which one specific MR method is most likely to yield highest statistical power for non-null associations, and lowest false discovery rates for null associations. This involves creating a 'gating function', whose purpose is to decide which expert to use for a specific MR analysis, given the parameter space that is occupied by that dataset. The metrics are a mixture of regression diagnostics and are described in Supplementary table 3.

The gating function needs to be trained using data for which the true causal effect is known, and to this end we generated a large number of simulated datasets. We trained the gating function using random forest learning algorithms that seek to identify sectors of the parameter space that are most likely to return accurate estimates for a particular expert. Figure 2a illustrates how the gating function is trained using simulated data. New datasets can then be applied to the trained mixture of experts (MoE) model (Figure 2b). We call this implementation MR-MoE 1.0.

It should be noted that in the original hierarchical mixture of experts approach of (27) the gating function and the experts share the same input space and are trained simultaneously using expectation-maximisation. In our approach the gating function is defined over a separate input space consisting of the parameter space of the experts, and is trained separately in a supervised setting using simulated data. As such our MR-MoE approach fits the framework of meta-learning (28–30) where machine learning methods are applied at the 'meta-level' to analyse the results of machine learning experiments at the 'base-level', in order to be able to understand and model the capabilities of each expert and recommend which expert should be applied to a given problem. From this perspective the 28 MR methods are the base-models, the gating function is the meta-model, and its meta-features are the parameters of the datasets.

Training and testing simulations

The MoE is trained using datasets generated from simulations (Figure 2a). A *dataset* is the minimum data required to perform 2SMR analysis - four columns comprising the SNP-exposure and SNP-outcome effects and standard errors, and rows corresponding to each SNP that is used as an instrument for the exposure. This *dataset* can be fed into any of the 28 experts to obtain MR causal effects. The simulations used to generate these datasets seek to cover a range of pleiotropic scenarios, including where some proportion of SNPs exhibit directional or balanced horizontal pleiotropy, or where SNPs influence confounding variables.

We simulate two individual level datasets for which there are N_x and N_y samples, and M SNPs, where each SNP has effect allele frequency of $p_m \sim U(0.05, 0.95)$. These datasets are used to obtain the SNP effects for the exposure trait x and the outcome trait y, respectively, using the following sampling criteria:

$$\begin{aligned}
N_x &= \{20000, ..., 500000\} \\
N_y &= \{20000, ..., 500000\} \\
K &= \{0, ..., 10\} \\
M_x &= \{1, ..., 200\} \\
M_y &= \{1, ..., 200\} \\
M_{u_k} &= \{5, ..., 30\}
\end{aligned}$$

The $M = M_x + M_y + \sum M_{u_k}$ SNPs can influence x directly, y directly, or some number of confounders u_k directly. Phenotypes for x and y are constructed using

$$x = \sum_{i}^{M_x} \beta_{gx,x,i} g_{x,i} + \sum_{j}^{M_y} \beta_{gy,x,j} g_{y,j} + \sum_{k}^{K} \beta_{ux,k} u_k + e_x$$

where $\beta_{gx,x}$ is the vector of effects of each of the M_x SNPs that influence x primarily, $\beta_{gy,x}$ is the vector of effects for the M_y SNPs on x, where the M_y SNPs influence y primarily but exhibit horizontal pleiotropic effects on x. We allow some proportion of these effects to be 0. β_{ux} is the vector of effects of each of the K confounders on x. Each u_k variable is constructed using

$$u = \sum_{l}^{M_u} \beta_{gu,l} g_l + e_l$$

and finally y is constructed using

$$y = \beta_{x,y}x + \sum_{i}^{M_y} \beta_{gy,y,j} g_{y,j} + \sum_{j}^{M_x} \beta_{gx,y,i} g_{x,i} + \sum_{k}^{K} \beta_{uy,k} u_k + e_y$$

where $\beta_{x,y}$ is the causal effect of x on y. We sample the distribution of direct SNP effects using

$$\begin{aligned}
\beta_{gx,x,i} &\sim N(0, \sigma^2_{gx,x}) \\
\sigma^2_{gx,x,i} &\sim U(0.01, 0.1) \\
\beta_{gy,y,j}\ N(0, \sigma^2_{gy,y}) \\
\sigma^2_{gy,y,j} &\sim U(0.01, 0.1)
\end{aligned}$$

Some proportion $floor(MS_x/M)$ of g_x SNPs and $floor(MS_y/M)$ of g_y SNPs, where s_x and $s_y \sim U(0,1)$, exhibit horizontal pleiotropy with effects sampled using

$$\begin{aligned}
\beta_{gx,y,i*} &\sim N(\mu_{gx,y}, \sigma^2_{gx,y}) \\
\mu_{gx,y,i*} &\sim U(-0.005, 0.005) \\
\sigma^2_{gx,y,i*} &\sim U(0.001, 0.01) \\
\beta_{gy,x,j*} &\sim N(\mu_{gy,x}, \sigma^2_{gy,x}) \\
\mu_{gy,x,j*} &\sim U(-0.005, 0.005) \\
\sigma^2_{gy,x,j*} &\sim U(0.001, 0.01)
\end{aligned}$$

The genetic influences on each of the confounders are sampled using

$$\begin{aligned}
\beta_{gu,u,l} &\sim N(0, \sigma^2_{gu,u}) \\
\sigma^2_{gu,u,l} &\sim U(0.01, 0.1)
\end{aligned}$$

The influence of each confounder on x and y is obtained using

$$\begin{aligned}
\beta_{u,x} &\sim N(0, \sigma^2_{u,x}) \\
\beta_{u,y} &\sim N(0, \sigma^2_{u,y})
\end{aligned}$$

Finally, 20% of the simulations have a null effect of $\beta_{x,y} = 0$, while the other remaining 80% have a true effect sampled from

$$\begin{aligned}
.\beta_{x,y} &\sim N(0, \sigma^2_{x,y}) \\
\sigma^2_{x,y} &\sim U(0.001, 0.1)
\end{aligned}$$

For each simulation we used linear regression to estimate the genetic effect of each SNP M on x in sample 1, and each SNP M on y in sample 2. We then perform MR analysis in both directions, mimicking GWAS by retaining SNPs that have $p < 5e-8$ in sample 1 to perform MR of x on y (the true causal direction for non-null simulations), and retaining SNPs that have $p < 5e-8$ in sample 2 to perform MR of y on x (the reverse causal direction for non-null simulations). We treat the summary data (effect sizes and standard errors) used for estimating $x \to y$ the summary data used for estimating $y \to x$ as two separate datasets Hence, for each simulation two datasets are generated which are analysed to produce 28 MR estimates each. We performed 100,000 simulations using these parameters, resulting in 200,000 datasets.

Optimisation function

We aim to maximise statistical power for datasets where $\beta_{x,y} \neq 0$ and minimise false discovery rates for datasets where $\beta_{x,y} = 0$. To train random forest decision trees to predict performance for a particular method $h(O_{w,d}, \mathbf{z}_d)$ is generated where the training set of input metrics for dataset d is $\mathbf{z}_d$ and the response (optimisation function) is

$$O_{w,d} = \begin{cases} 1, & \text{if } \beta_{x,y} \neq 0 \text{ and } p_{m,d} < 0.01 \\ 1, & \text{if } \beta_{x,y} = 0 \text{ and } p_{m,d} > 0.1 \\ 0, & \text{otherwise} \end{cases}$$

where $p_{m,d}$ is the p-value for method m on dataset d.

Strategy

For each training dataset we record a set of 53 metrics $\mathbf{z}_d$ (Supplementary table 3), and an outcome $O_{w,d}$, which is a measure of how well that method performed for each particular simulated dataset. For each of our 28 methods, we need to create a model that predicts the performance of the method based on metrics generated from a dataset. To do this, for each method we train a random decision forest to predict that method's performance using the dataset's metrics. The random forest approach is well suited to this problem because there are likely to be non-continuous combinations of different metrics that improve on prediction over, for example, a simple linear model that does not learn about interactions.

As a simple hypothetical example - if a dataset has a single outlier but is otherwise exhibiting no heterogeneity then the following methods could arguably perform well:

- an IVW fixed effects analysis with the outlier removed, should the Steiger test be able to detect the outlier
- a median based approach
- a mode based approach

deciding between these methods requires finding, in general, which will minimise false discovery rates and maximise true positive rates for that particular scenario. In this example the IVW with Steiger filtering would likely be the clear winner because of its superior statistical power. Countless other scenarios could arise. For example if 100% of instruments are invalid but the InSIDE assumptio is met then the MR-Egger is likely to be most effective; if 40% of instruments are invalid then a median-based approach is likely most effective; and if 80% of instruments are invalid then a mode-based approach is likely most effective.

Having generated random forest decision trees for each of the 28 methods using 133,000 of the simulations, we then applied them to the remaining 67,000 datasets to predict which method would have the highest performance for each of the remaining datasets. Finally we compare the performance of the method selected by the MoE against all remaining methods. The default settings for the `randomForest` package in R (31) were used to train the models. MR-MoE 1.0 is implemented in the TwoSampleMR R package available at github.com/MRCIEU/TwoSampleMR (11).

Graph database of MR estimates

The set of MR estimates obtained from this analysis are recorded in a Neo4j graph database. Because each association has up to 28 different estimates, for simplicity we distill this down to a single 'best estimate' for each association using the following rules:

1. If the number of variants after Steiger filtering is greater than 5 then apply the MoE to obtain the best method
2. If the number of variants after Steiger filtering is less than or equal to 5 but greater than 1 then use the IVW random effects approach on the filtered set of variants
3. If there is 1 variant retained in the Steiger approach then use the Wald ratio on the remaining variant
4. If there are no variants remaining after Steiger filtering then declare no causal association.

For specific hypotheses we strongly recommend that estimates from all sensitivity analyses are scrutinised and reported. The graph can be queried directly using the cypher language at http://eve-neo4j.mrbase.org or through a basic web interface at http://eve.mrbase.org.

Results

Steiger filtering improves reliability

As statistical power for GWAS studies improves the likelihood of a significant association being discovered that doesn't act primarily on the trait of interest increases (Figure 1a). We evaluated the efficacy of the Steiger filtering approach for improving instrument selection using 100,000 simulated datasets comprising both null and non-null causal models. For each dataset, instruments were selected based on the tophits strategy and the Steiger filtering strategy, and MR was performed using 14 different methods based on instruments selected from each of these strategies.

Figure 1b shows that the tophits strategy led to over half of the instruments being primarily associated with either confounders or the outcome phenotype, not the exposure phenotype. The proportion of invalid instruments due to reverse cause increased as GWAS discovery sample size increased. Applying Steiger filtering reduced this to 25%. Consequently, the false discovery rates (FDR) for 12 of the 14 methods reduced substantially when applied using Steiger filtered instruments (Figure 1c). The true positive rates for the methods based on Steiger filtering did however reduce slightly for 10 of the 14 methods.

Mixture of experts method selection improves over any single method

Following evidence that the Steiger filtering approach can improve on existing methods, we next hypothesised that a mixture of experts (MoE) model would be able to predict the most appropriate of the 28 MR methods and instrument selection strategies to apply to a particular dataset based on its characteristics (Figure 2).

The ability to predict the performance for each of these methods is shown in Supplementary table 2. The prediction R^2 of whether a dataset's status was truly null ($\beta_{xy} = 0$) or non-null ($\beta_{xy} \neq 0$) against the method's prediction of the dataset's status, ranged between 0.04 and 0.24. The method performance prediction was most effective for the Egger random effects model with Steiger filtering. The dataset characteristics with the most importance for each of the predictors differed substantially between each dataset, as well as the frequency for which each of the methods was selected in the testing datasets. The FDR of each method when chosen, compared to their averages across all datasets, reduced; and likewise the true positive rates of each method when chosen increased compared to their simulation-wide averages.

We compared the MoE performance in the simulations against each of the 28 methods, testing to see if it outperformed all other single methods. Figures 3a and 3b show that the MoE approach had the best general performance, obtaining an AUC of 0.84 in terms of classifying the simulations as being null or non-null. Notably, the next best methods were median and mode based estimators using Steiger filtering. Under the assumption of widespread pleiotropy it is clear that all methods suffer from high false discovery rates.

Automated MR analysis of 2407 phenotypes

We applied our analysis using summary data for 2407 phenotypes, including 149 complex traits and diseases, 575 metabolites (32,33) and 1683 plasma protein levels (34). For the protein levels only the instruments were available, so they could only be evaluated as exposure phenotypes. The complex traits and diseases and metabolite levels had complete summary data available in MR-Base, thus they be evaluated as both exposures (if they had significant instruments) and outcomes. Together, we evaluated 715681 relationships. The majority of these associations could only be evaluated using fewer than 5 SNPs, and so the Wald ratio or IVW fixed effects methods were used, but for 61029 associations the MR-MoE approach was applied.

There were 5660 associations following Bonferroni correction ($p < 7.0 \times 10^{-8}$). Of these 2918 were obtained from the MR-MoE analysis, while the remainder were estimated using fewer than 5 SNPs using the Wald ratio or IVW fixed effects methods.

The frequencies of the methods chosen by the MR-MoE analysis are shown in Supplementary table 3. Amongst those deemed 'significant', the IVW fixed effects analysis using tophits instruments (the only method applicable when there is no evidence of horizontal pleiotropy of any sort) was selected in only 10.4% of cases. This indicates that horizontal pleiotropy is likely to be pervasive.

Putative associations involving years of schooling

As an example of the database we performed a look up of associations where LDL cholesterol as measured in the GLGC consortium (35) had a false discovery rate of 0.05 using the MoE approach. This returned 287 putative associations, amongst which 111 involved LDL cholesterol influencing other traits and 176 involved other traits influencing LDL cholesterol. A large proportion of these traits were metabolites, which are dominated by lipid fractions. Filtering to exclude the metabolomic studies ((32,33)) returned 27 associations with 23 traits (Figure 4a). There was a strong association with coronary heart disease (0.41 log(OR) per SD, 0.31-0.52, $p = 2.7 \times 10^{-12}$). Here the IVW random effects method was chosen after Steiger filtering, indicating a prediction of HP being present amongst the instruments.

We also performed a look up of associations that causally influence years of schooling (36), or that years of schooling itself causally influences. Using a false-discovery rate of 0.05, 45 traits were returned as having some direction of causality with years of schooling (Figure 4b). Several of these putative relationships appear to be plausible, for example the influence of years of schooling on lower risk of Alzheimer's disease has recently been examined (37). However it is clear that there are serious limitations in relying on MR as a panacea for causal inference. For example, bi-directional causal relationships with college completion exist because of the similarity in trait definitions; and there are several traits (e.g. childhood intelligence and birth weight) which appear to be causally downstream of years of schooling despite this being temporally impossible. Though explanations could be conjured for this association, for example parents' schooling influences childhood intelligence, or the genetic instruments are shared across the two traits, an intuitive interpretation of causality cannot be applied.

The facility to perform these scans has been automated in the MR of everything vs everything (MR-EvE) web application, http://eve.mrbase.org.

Discussion

The trend of increasing GWAS sample sizes continues, but while the opportunity that this affords MR to be furnished with more instruments is typically welcomed, here we have shown that it is invalid instruments, exhibiting horizontal pleiotropy or reverse causation, that are more likely to be identified than valid ones. Strategising how MR is to be used in practice, therefore, must consider horizontal pleiotropy to be the rule rather than the exception. Following a recent flurry in development of MR methodology, we have devised a machine learning approach, MR-MoE, that seeks to predict the performance of each MR method, selecting the one which is most likely to maximise power whilst minimising FDR for a specific dataset. The mixture of experts approach that we present here is trained using random forest decision trees applied to extensive and diverse simulations, and we demonstrate that it makes substantial performance improvements over any other single method in the presence or absence of extensive horizontal pleiotropy. As such, our method contributes to the field of meta-learning but also to the rapidly growing field of automatic model selection and hyper-parameter optimization in machine learning (38).

Crucially however, we observed in our simulations that none of the available methods were conservative and all, including MR-MoE, had FDR above 0.2. Instrumental variables in MR are typically chosen blind, with GWAS significance being the only criteria. We illustrated how confounders and reverse causal associations can easily lead to invalid instruments being used in MR, and that even if those instruments that only influence the exposure directly are used, most randomly generated patterns of horizontal pleiotropy cannot be adequately accounted for by any single MR method.

We applied the method to the curated set of GWAS summary data present in MR-Base (11), generating several million MR estimates. MR-MoE selected a method that indicated some pattern of horizontal pleiotropy for almost 90% of cases, reinforcing the notion that it is the rule rather than the exception. Based on our simulations we expect that those associations that we reported to be 'significant' are liable to high type 1 error rates.

In addition to correctly handling horizontal pleiotropy, there are many other limitations that prevent MR from being a panacea for causal inference. Many of the associations are biologically impossible, for example where early life phenotypes appear to be influenced by later stage phenotypes. Though these associations can be informative, their interpretations as causal relationships are far from clear. Similarly, often disease traits appear to causally relate to other phenotypes, but GWAS is typically performed on the liability scale, hence the causal estimate reflects not the presence or absence of disease, but the underlying risk of disease. Again, interpretation of such associations can be problematic.

A large proportion of the associations that were estimated used only a single instrumental variable. Methods are emerging to attempt to delineate between models of reverse cause, pleiotropy and multiple causal variants in the same region (39,40). Separating vertical from horizontal pleiotropy with a single instrument is not yet possible in the two-sample MR framework though the use of mediation-based approaches may be informative in the one-sample setting (7,41). Other problems can also manifest, for example frailty effects could induce associations for late-onset traits (42); genetic variants could strongly relate to several phenotypes making it difficult to ascertain which of them is the causal exposure (43); and the measured feature for which genetic associations are known could relate in complicated ways to the biological entity that are truly causal (2).

We also encounter a potential new problem in using machine learning approaches to infer the correct model for a particular dataset, namely, that the data is choosing the model. This is liable to contribute to elevated type 1 error rates, even though we have attempted to separate the information used for optimisation from the information used to predict performance. The extent to which a machine based method selector suffers from this more than a human manually inferring the result most likely to be true is not clear. Though MR-MoE does exhibit higher type 1 error rates than are desirable, they still remain amongst the lowest compared to all other MR methods that do not suffer from this issue.

The field of MR is evolving. The advent of GWAS databases (9,44) and automated 2SMR (11) has trivialised the analytical aspect of investigating specific causal hypotheses. Despite the limitations described above, the construction of a causal graph of 'everything versus everything' does have appeal. First, though causality is not guaranteed by MR, it can still be highly informative for confirming or negating specific hypotheses. Second, it paves the way for new approaches to search for novel putative associations and improve reliability in single analyses. The potential to exploit GWAS summary data within the properties of graph databases to aid with both of these endeavours could be transformative in biological research.

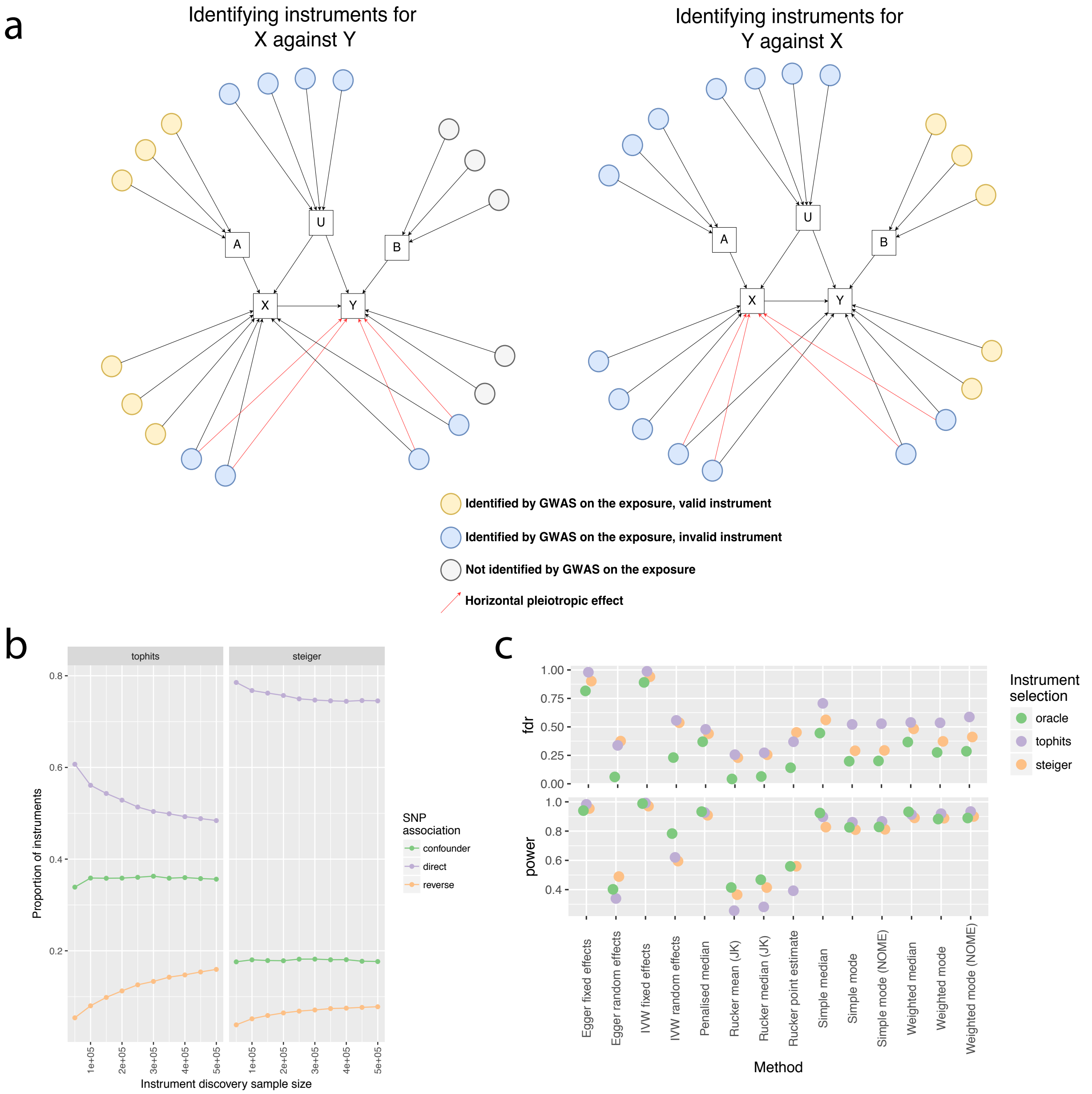
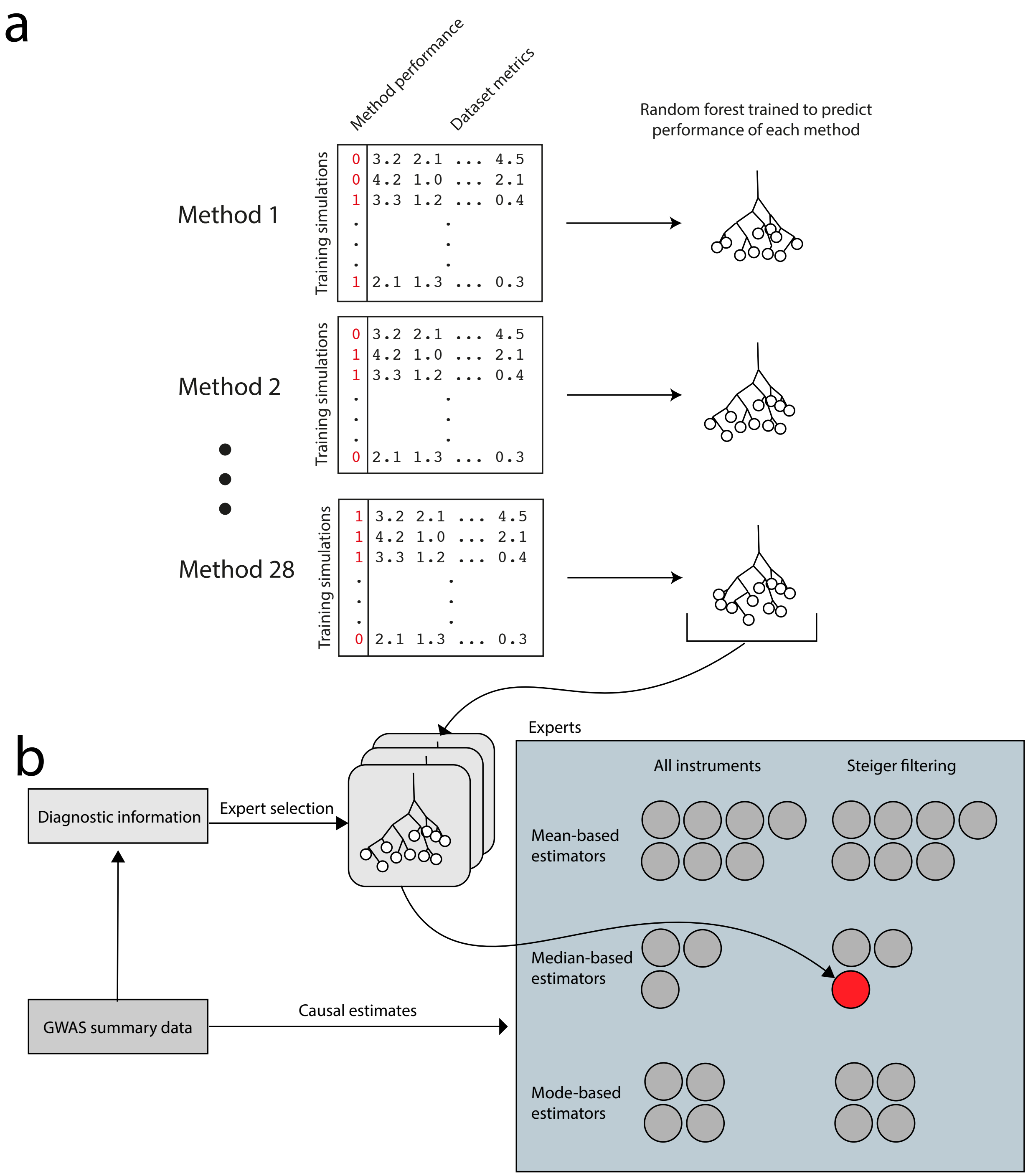


Figure 1: Simulations. a) Schematic of how GWAS with sufficient power can lead to the selection of instrumental variables that are invalid. We used arbitrary numbers of SNPs and confounders to simulate GWAS summary datasets. b) Any SNP that has a direct influence on the exposure, or an influence on a non-confounding intermediate variable, is considered a 'direct' effect. The y-axis shows the proportion of instruments selected for analysis due that are either direct associations with the exposure, instruments for the outcome (reverse), or instruments for confounding traits (confounder). The proportions are compared over a range of different exposure discovery sample sizes (x-axis) and using either the tophits approach (left) or the Steiger approach (right) for instrument selection. c) Top: The false discovery rates from null simulations for each of the 14 methods using either tophits, Steiger filtered variants, or variants that are known to be directly associated with the exposure (oracle, note that direct effects can still exhibit horizontal pleiotropy in these simulations). Bottom: The statistical power to detect true causal associations in the non-null simulations.



Method 1

```
0 | 3.2  2.1  ...  4.5
0 | 4.2  1.0  ...  2.1
1 | 3.3  1.2  ...  0.4
.           .
.           .
.           .
1 | 2.1  1.3  ...  0.3
```

Method 2

```
0 | 3.2  2.1  ...  4.5
1 | 4.2  1.0  ...  2.1
1 | 3.3  1.2  ...  0.4
.           .
.           .
.           .
0 | 2.1  1.3  ...  0.3
```

Method 28

```
1 | 3.2  2.1  ...  4.5
1 | 4.2  1.0  ...  2.1
1 | 3.3  1.2  ...  0.4
.           .
.           .
.           .
0 | 2.1  1.3  ...  0.3
```

Figure 2: Mixture of experts. a) Training. Datasets are simulated that have either null or non-null causal relationships, and MR is performed by each of the 28 available MR methods. In the toy datasets, the columns in black represent 53 metrics about each of the 67,000 training simulations. The columns in red are specific to each method, they represent how well that method performed in obtaining the correct answer for each of the datasets. Random forests are used to learn the parameter space of the 53 metrics in which a particular dataset is likely to perform well. Together, this creates 28 random forest decision trees, one for each method. b) Application. For a GWAS summary dataset, our objective is to choose the method most likely to return the correct causal estimate. Metrics are generated from the dataset and fed into each of the 28 random forest

decision trees. This provides us with 28 performance predictions. Finally, we use the method for which the performance prediction is highest.

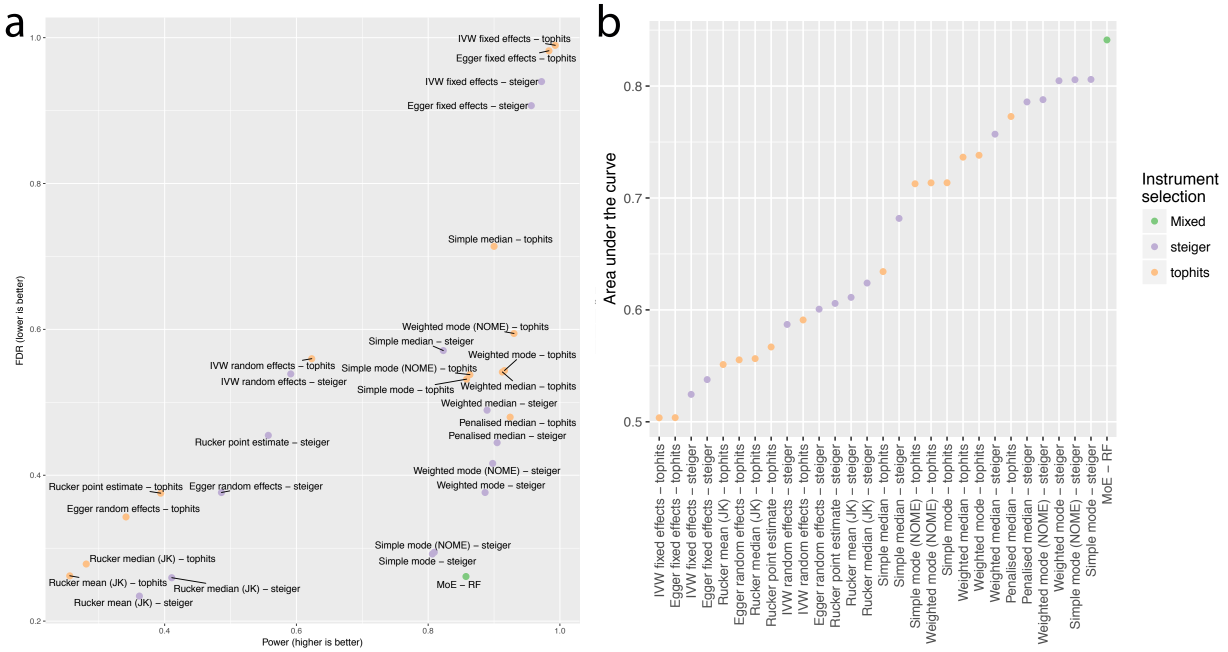


Figure 3: Performance of MoE against all other methods. a) The power for non-null datasets is plotted against the FDR for null datasets for each of the 28 methods, plus MoE. No single method achieved nominal FDR for these simulations. b) Calculating the area under the ROC curve from the values in (a) we plotted the performance in order from lowest to highest. Under the assumption of pervasive horizontal pleiotropy, the MoE approach is likely most effective than any other single method.

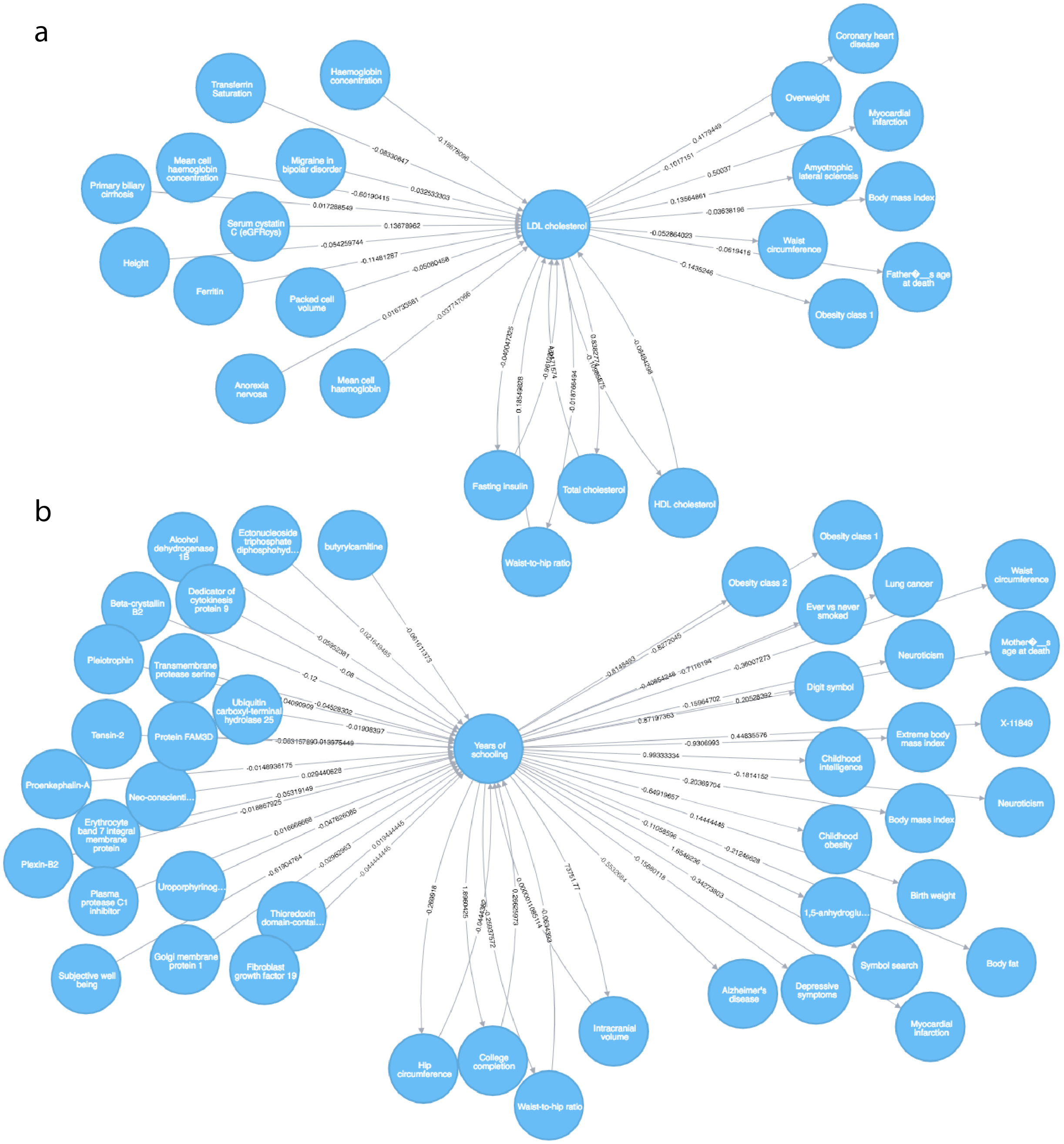


Figure 4: Lookup of causal associations involving a) LDL cholesterol and b) Years of Schooling. In a) the nodes are filtered to not include those obtained from the Shin et al or Kettunen et al metabolomic studies. The arrows denote causal direction and the values on the arrows denote the causal effect estimate. Only those relationships are shown for which the $FDR < 0.05$.

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology [Internet]. 2003 Feb;32(1):1–22. Available from: http://www.ije.oxfordjournals.org/cgi/doi/10.1093/ije/dyg070

2. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics. 2014 Jul;23(R1):R89——R98.

3. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015;44(2):512–25.

4. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology [Internet]. 2016 May;40(4):304–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27061298 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4849733 http://doi.wiley.com/10.1002/gepi.21965

5. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine [Internet]. 2017; Available from: http://doi.wiley.com/10.1002/sim.7221

6. Hartwig FP, Davey Smith G, Bowden J. Robust Inference In Two-Sample Mendelian Randomisation Via The Zero Modal Pleiotropy Assumption. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/04/10/126102

7. Hemani G, Tilling K, Davey Smith G. Orienting The Causal Relationship Between Imprecisely Measured Traits Using Genetic Instruments. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/03/15/117101

8. Verbanck M, Chen C-Y, Neale B, Do R. Widespread pleiotropy confounds causal relationships between complex traits and diseases inferred from Mendelian randomization. bioRxiv [Internet]. 2017; Available from: http://www.biorxiv.org/content/early/2017/06/30/157552

9. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies, available at http://www.genome.gov/gwastudies. Accessed 12/10/2010. 2010.

10. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. American journal of epidemiology [Internet]. 2013 Oct;178(7):1177–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3783091{\&}tool=pmcentrez{\&}rendertype=abstract

11. Hemani G, Zheng J, Wade KH, Laurin C, Elsworth B, Burgess S, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. BioRxiv. 2016;10.1101/07.

12. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. bioRxiv [Internet]. 2017; Available from: http://www.biorxiv.org/content/early/2017/07/26/168674

13. Wright S. Evolution and the Genetics of Populations. In: Evolution and the genetics of populations. Chicago: University of Chicago Press; 1968.

14. Wagner GP, Zhang J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nature Reviews Genetics [Internet]. 2011 Mar;12(3):204–13. Available from: http://www.nature.com/doifinder/10.1038/nrg2949

15. Hill WG, Zhang X-S. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. Nature Reviews Genetics [Internet]. 2012

Feb;13(4):296. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22349131

16. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. International Journal of Epidemiology [Internet]. 2016 Dec;45(6):1717–26. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyx028

17. Xu L, Lin SL, Schooling CM. A Mendelian randomization study of the effect of calcium on coronary artery disease, myocardial infarction and their risk factors. Scientific Reports [Internet]. 2017 Feb;7:42691. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28195141 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5307362 http://www.nature.com/articles/srep42691

18. Larsson SC, Burgess S, Michaëlsson K, TB H, WH C, Y P. Association of Genetic Variants Related to Serum Calcium Levels With Coronary Artery Disease and Myocardial Infarction. JAMA [Internet]. 2017 Jul;318(4):371. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.8981

19. Prins BP, Abbasi A, Wong A, Vaez A, Nolte I, Franceschini N, et al. Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study. Hay PJ, editor. PLOS Medicine [Internet]. 2016 Jun;13(6):e1001976. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27327646 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4915710 http://dx.plos.org/10.1371/journal.pmed.1001976

20. Inoshita M, Numata S, Tajima A, Kinoshita M, Umehara H, Nakataki M, et al. A significant causal association between C-reactive protein levels and schizophrenia. Scientific Reports [Internet]. 2016 Sep;6(1):26105. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27193331 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4872134 http://www.nature.com/articles/srep26105

21. Rucker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics [Internet]. 2011 Jan;12(1):122–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20656692 https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxq046

22. Han C. Detecting invalid instruments using L 1-GMM. Economics Letters. 2008;101(3):285–7.

23. Kang H, Zhang A, Cai TT, Small DS. Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization. Journal of the American Statistical Association. 2016;111(513):132–44.

24. Steiger JH. Tests for comparing elements of a correlation matrix. Psychological Bulletin. 1980;87(2):245–51.

25. Corbin LJ, Richmond RC, Wade KH, Burgess S, Bowden J, Smith GD, et al. BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding Causal Estimates Using Mendelian Randomization. Diabetes [Internet]. 2016 Oct;65(10):3002–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27402723 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5279886

26. Lee SH, Wray NR. Novel genetic analysis for case-control genome-wide association studies: quantification of power and genomic prediction accuracy. PLoS One. 2013;8(8):e71494.

27. Jordan MI, Jacobs RA. Hierarchical Mixtures of Experts and the EM Algorithm. Neural Computation [Internet]. 1994 Mar;6(2):181–214. Available from: http://www.mitpressjournals.org/doi/10.1162/neco.1994.6.2.181

28. Brazdil P, Carrier CG, Soares C, Vilalta R. Metalearning: Applications to data mining. Springer Science & Business Media; 2008.

29. Smith-Miles KA. Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM

Computing Surveys (CSUR). 2009;41(1):6.

30. Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies. Artificial intelligence review. 2015;44(1):117–30.

31. Liaw A, Wiener M. Classification and Regression by randomForest. R News [Internet]. 2002;2(3):18–22. Available from: http://cran.r-project.org/doc/Rnews/

32. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nature genetics [Internet]. 2014 Jun;46(6):543–50. Available from: http://dx.doi.org/10.1038/ng.2982

33. Kettunen J, Demirkan A, Wurtz P, Draisma HHM, Haller T, Rawal R, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun [Internet]. 2016 Mar;7. Available from: http://dx.doi.org/10.1038/ncomms11122 http://10.0.4.14/ncomms11122

34. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Consequences Of Natural Perturbations In The Human Plasma Proteome. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/05/05/134551

35. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nature Genetics [Internet]. 2013;45(11):1274–83. Available from: http://www.nature.com/doifinder/10.1038/ng.2797

36. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature [Internet]. 2016 May;533(7604):539–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27225129 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4883595 http://www.nature.com/doifinder/10.1038/nature17671

37. Anderson E, Wade KH, Hemani G, Bowden J, Korologou-Linden R, Davey Smith G, et al. The Causal Effect Of Educational Attainment On Alzheimer's Disease: A Two-Sample Mendelian Randomization Study. bioRxiv [Internet]. 2017; Available from: http://www.biorxiv.org/content/early/2017/04/17/127993

38. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research. 2016;17:1–5.

39. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics [Internet]. 2016 Mar;48(5):481–7. Available from: http://www.nature.com/doifinder/10.1038/ng.3538

40. Richardson TG, Zheng J, Davey Smith G, Timpson NJ, Gaunt TR, Relton CL, et al. Causal epigenome-wide association study identifies CpG sites that influence cardiovascular disease risk. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/04/29/132019

41. Wang L, Michoel T. Efficient And Accurate Causal Inference With Hidden Confounders From Genome-Transcriptome Variation Data. bioRxiv [Internet]. 2017; Available from: http://www.biorxiv.org/content/early/2017/04/19/128496

42. Noyce AJ, Kia DA, Hemani G, Nicolas A, Price TR, De Pablo-Fernandez E, et al. Estimating the causal influence of body mass index on risk of Parkinson disease: A Mendelian randomisation study. Brayne C, editor. PLOS Medicine [Internet]. 2017 Jun;14(6):e1002314. Available from: http://dx.plos.org/10.1371/journal.pmed.1002314

43. Burgess S, Freitag DF, Khan H, Gorman DN, Thompson SG. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. PloS one [Internet]. 2014 Jan;9(10):e108891. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0108891

44. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype–phenotype associations. Bioinformatics [Internet]. 2016 Oct;32(20):3207–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27318201

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5048068
https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw373