Selecting causal risk factors from high-throughput experiments using multivariable Mendelian randomization

Verena Zuber[1,2], Johanna Maria Colijn[3,4], Caroline Klaver[3,4,5], and Stephen Burgess[1,6]

[1]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, UK
[2]Department of Epidemiology and Biostatistics, Imperial College London, UK
[3]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
[4]Department of Ophthalmology, Erasmus University Medical Center, Rotterdam, The Netherlands
[5]Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands
[6]MRC/BHF Cardiovascular Epidemiology Unit, School of Clinical Medicine, University of Cambridge, UK

May 2, 2019

Abstract

Modern high-throughput experiments provide a rich resource to investigate causal determinants of disease risk. Mendelian randomization (MR) is the use of genetic variants as instrumental variables to infer the causal effect of a specific risk factor on an outcome. Multivariable MR is an extension of the standard MR framework to consider multiple potential risk factors in a single model. However, current implementations of multivariable MR use standard linear regression and hence perform poorly with many risk factors.

Here, we propose a novel approach to two-sample multivariable MR based on Bayesian model averaging (MR-BMA) that scales to high-throughput experiments. In a realistic simulation study, we show that MR-BMA can detect true causal risk factors even when the candidate risk factors are highly correlated. We illustrate MR-BMA by analysing publicly-available summarized data on metabolites to prioritise likely causal biomarkers for age-related macular degeneration.

Wordcount: 142/150

Mendelian randomization (MR) is the use of genetic variants to infer the presence or absence of a causal effect of a risk factor on an outcome. Under the assumption that the genetic variants are valid instrumental variables, this causal effect can be consistently inferred even in the presence of unobserved confounding factors [1]. The instrumental variable assumptions are illustrated by a directed acyclic graph as shown in Figure 1 [2].

Recent years have seen an explosion in the size and scale of datasets with biomarker data from high-throughput experiments and concomitant genetic data. These biomarkers include proteins [3], blood cell traits [4], metabo-

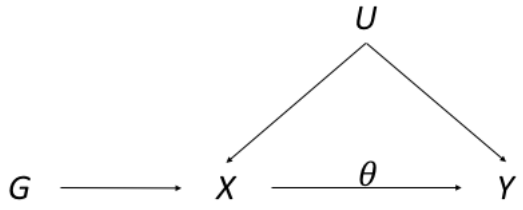

Figure 1: Directed acyclic graph of instrumental variable assumptions made in univariable Mendelian randomization. G = genetic variant(s), X = risk factor, Y = outcome, U = confounders, θ = causal effect of interest.

lites [5] or imaging phenotypes such as from cardiac image analysis [6]. High-throughput experiments provide ideal data resources for conducting MR investigations in conjunction with case-control datasets providing genetic associations with disease outcomes (such as from CARDIoGRAMplusC4D for coronary artery disease [7], DIAGRAM for type 2 diabetes [8], or the International Age-related Macular Degeneration Genomics Consortium [IAMDGC] for age-related macular degeneration [9]). In addition to their untargeted scope, one specific feature of high-throughput experiments is a distinctive correlation pattern between the candidate risk factors shaped by latent biological processes.

Multivariable MR is an extension of standard (univariable) MR that allows multiple risk factors to be modelled at once [10]. Whereas univariable MR makes the assumption that genetic variants specifically influence a single risk factor, multivariable MR makes the assumption that genetic variants influence a set of multiple measured risk factors and thus accounts for measured pleiotropy. Our aim is to use genetic variation in a multivariable

MR paradigm to select which risk factors from a set of related and potentially highly correlated candidate risk factors are causal determinants of an outcome. Existing methods for multivariable MR are designed for a small number of risk factors and do not scale to the dimension of high-throughput experiments. We therefore seek to develop a method for multivariable MR that can select and prioritize biomarkers from high-throughput experiments as risk factors for the outcome of interest. In this context we propose a Bayesian model averaging approach (MR-BMA) that scales to the dimension of high-throughput experiments and enables risk factor selection from a large number of candidate risk factors. MR-BMA is formulated on two-sample summarized genetic data which is publicly available and allows the sample size to be maximized.

To illustrate our approach, we analyse publicly available summarized data from a metabolite genome-wide association study (GWAS) on nearly 25 000 participants to rank and prioritise metabolites as potential biomarkers for age-related macular degeneration. Data are available on genetic associations with 118 circulating metabolites measured by nuclear magnetic resonance (NMR) spectroscopy [11] from `http://computationalmedicine.fi/data#NMR_GWAS`. This NMR platform provides a detailed characterisation of lipid subfractions, including 14 size categories of lipoprotein particles ranging from extra small (XS) high density lipoprotein (HDL) to extra-extra-large (XXL) very low density lipoprotein (VLDL). For each lipoprotein category, measures are available of total cholesterol, triglycerides, phospholipids, and cholesterol esters, and additionally the average diameter of the lipoprotein particles. Apart from lipoprotein measurements, this metabolite GWAS estimated ge-

netic associations with amino acids, apolipoproteins, fatty and fluid acids, ketone bodies, and glycerides. We assess the performance of our proposed method in a simulation study with scenarios motivated by the metabolite GWAS and by publicly available summary data on blood cell traits measured on nearly 175 000 participants [4].

Results

Multivariable Mendelian randomization and risk factor selection

Multivariable MR is an extension of the standard MR paradigm (Figure 1) to model not one, but multiple risk factors (Figure 2), thus accounting for measured pleiotropy. We consider a two-sample framework, where the genetic associations with the outcome (sample 1) are regressed on the genetic associations with all the risk factors (sample 2) in a multivariable regression which is implemented in an inverse-variance weighted (IVW) linear regression. Each genetic variant contributes one data point (or observation) to the regression model. Weights in this regression model are proportional to the inverse of the variance of the genetic association with the outcome. This is to ensure that genetic variants having more precise association estimates receive more weight in the analysis. The causal effect estimates from multivariable MR represent the direct causal effects of the risk factors in turn on the outcome when all the other risk factors in the model are held constant [12, 13] and Supplementary Figure 1). Including multiple risk factors into a single

model allows genetic variants to have pleiotropic effects on the risk factors in the model referred to as "measured pleiotropy" [14].

However, the current implementation of multivariable MR is not designed to consider a high-dimensional set of risk factors and is not suitable to select biomarkers from high-throughput experiments.

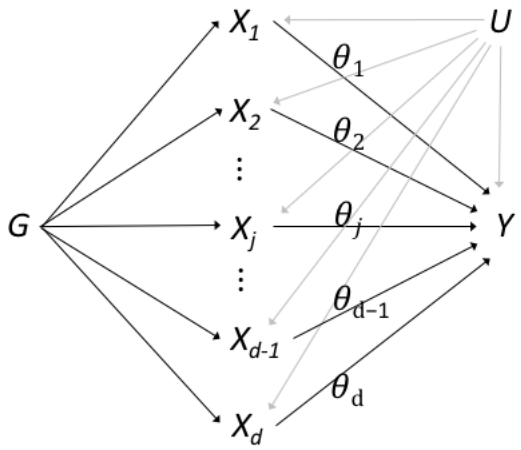

Figure 2: Directed acyclic graph of instrumental variable assumptions made in multivariable Mendelian randomization. G = genetic variants, X_j = risk factor j for $j = 1, \ldots, d$, Y = outcome, U = confounders, θ_j = causal effect of risk factor j.

To allow joint analysis of biomarkers from high-throughput experiments in multivariable MR, we cast risk factor selection as variable selection in the same weighted linear regression model as in the IVW method. Formulated in a Bayesian framework (for full details we refer to the Methods section) we use independence priors and closed-form Bayes factors to evaluate the posterior probability (PP) of specific models (i.e. one risk factor or a combination of multiple risk factors). In high-dimensional variable selection, the evidence

for one particular model can be small because the model space is very large and many models might have comparable evidence. This is why MR-BMA uses Bayesian model averaging (BMA) and computes for each risk factor its marginal inclusion probability (*MIP*), which is defined as the sum of the posterior probabilities over all models where the risk factor is present. MR-BMA reports the model averaged causal effects (*MACE*), representing the direct causal effect of a risk factor on the outcome averaged across these models. As we show in a simulation study based on real biomarker data, MR-BMA enables sparse modelling and hence a better and more stable detection of the true causal risk factors than either the conventional IVW method or other variable selection methods.

Detection of invalid and influential instruments

Invalid instruments may be detected as outliers with respect to the fit of the linear model. Outliers may arise for a number of reasons, but they are likely to arise if a genetic variant has an effect on the outcome that is not mediated by one or other of the risk factors – an unmeasured pleiotropic effect. To quantify outliers we use the Q-statistic, which is an established tool for identifying heterogeneity in meta-analysis [15]. More precisely, to pinpoint specific genetic variants as outliers we use the contribution q of the variant to the overall Q-statistic, where q is defined as the weighted squared difference between the observed and predicted association with the outcome.

Even if there are no outliers, it is advisable to check for influential observations and re-run the approach omitting that influential variant from the

analysis. If a particular genetic variant has a strong association with the outcome, then it may have undue influence on the variable selection, leading to a model that fits that particular observation well, but other observations poorly. To quantify influential observations, we suggest to use Cook's distance (Cd) [16]. We illustrate the detection of influential points and outliers in the applied example and provide more details in the Methods.

Simulation results

To assess the performance of the proposed method, we perform a simulation study in three scenarios based on real high-dimensional data. We compare the performance of the conventional approach (Multivariable IVW regression), the Lars [17], Lasso, and Elastic Net [18] penalised regression methods developed for high-dimensional regression models, our novel MR-BMA method, and the model with the highest posterior probability from the BMA procedure (best model). We seek to evaluate two aspects of the methods: 1) how well can the methods select the true causal risk factors, and 2) how well can the methods estimate causal effects. Risk factor selection is evaluated using the receiver operating characteristic (ROC) curve, where the true positive rate is plotted against the false positive rate. True positives are defined as the risk factors in the generation model that have a non-zero causal effect. Causal estimation is evaluated by calculating the mean squared error (MSE) of estimates, which is defined as the squared difference between the estimated causal effect and the true causal effect. The MSE of an estimator decomposes into the sum of its squared bias and its variance.

Genetic associations with the risk factors are obtained from three different scenarios. Two scenarios are based on the NMR metabolite GWAS by [11], where we use as instrumental variables $n = 150$ independent genetic variants that were associated with any of three composite lipid measurements (LDL-cholesterol, triglycerides, or HDL-cholesterol) at a genome-wide level of significance ($p < 5 \times 10^{-8}$) in a large meta-analysis of the Global Lipids Genetics Consortium [19] . In Scenario 1, we consider a small set of $d = 12$ randomly selected risk factors, and in Scenario 2 a larger set of $d = 92$ risk factors (Supplementary Figure 2). Scenario 3 is based on publicly available summary data on $d = 33$ blood cell traits measured on nearly 175 000 participants [4]. Using all genetic variants that were genome-wide significant for any blood cell trait, we have $n = 2667$ genetic variants as instrumental variables. For each scenario, we generate the genetic associations for the outcome based on four random risk factors having a positive effect in Setting A and on eight random risk factors, of which four have a positive and four have a negative effect, in Setting B. Additionally, we vary the proportion of variance in the outcome explained by the causal risk factors. Each simulation setting is repeated 1000 times. Full detail of the generation of the simulated outcomes is given in the Supplementary Methods.

Looking at a small set of $d = 12$ risk factors in the NMR metabolite data of which four risk factors are true causal ones (Scenario 1, Setting A), we see that MR-BMA is dominating all other methods in terms of area under the ROC curve (see Figure 3 A). Next best methods are Lasso, Elastic Net, the Bayesian best model, and Lars. The standard IVW method gives the worst performance. Similar results were obtained when varying the variance

in the outcome explained by the risk factors (Setting A in Supplementary Figure 3 and Setting B in Supplementary Figure 4). With respect to the MSE of estimates (Table 1), MR-BMA has the lowest MSE in almost all scenarios followed by Elastic Net, Lasso, the Bayesian best model, and then Lars. Elastic Net has the lowest MSE for $R^2 = 0.5$ in setting B. The highest MSE is seen for the IVW method, which provides unbiased estimates, as can be seen in Supplementary Figure 5 and 6, but at the price of a high variance.

When increasing the number of risk factors to $d = 92$ while keeping the number of true causal risk factors constant to four (Scenario 2, Setting A), the standard IVW method fails to distinguish between true causal and false causal risk factors and provides a ranking of risk factors which is nearly random as shown in the ROC curve in Figure 3 B) and Supplementary Figures 7 and 8. Despite being unbiased (see Supplementary Figures 9 and 10), the variance of the IVW estimates is large and prohibits better performance. In contrast, Lars, Lasso, Elastic Net, and MR-BMA provide causal estimates which are biased towards zero, but have much reduced variance compared to the IVW estimates. The Lasso provides sparse solutions with many of the causal estimates set to zero. This allows the Lasso and Elastic Net to have relatively good performance at the beginning of the ROC curve, but their performance weakens when considering more risk factors. The best performance is terms of the ROC characteristics is observed for MR-BMA. In terms of MSE (Table 1), the dominant role of the variance of the IVW estimate becomes again apparent as the IVW method has a thousand times larger MSE than MR-BMA, which has the lowest MSE for all scenarios considered.

In the blood cell trait data (Scenario 3), MR-BMA has again the lowest

MSE, followed by the regularised regression approaches and the best model in the Bayesian approach. Despite a large sample size ($n = 2667$) and comparatively low dimension of the risk factor space ($d = 33$), the IVW approach is the only unbiased method at the cost of an inferior detection of true positive risk factors (Supplementary Figures 12 and 13) and a large variance (Supplementary Figures 14 and 15), and consequently a MSE which is in a magnitude of a hundred larger than other methods designed for high-dimensional data analysis (Table 1).

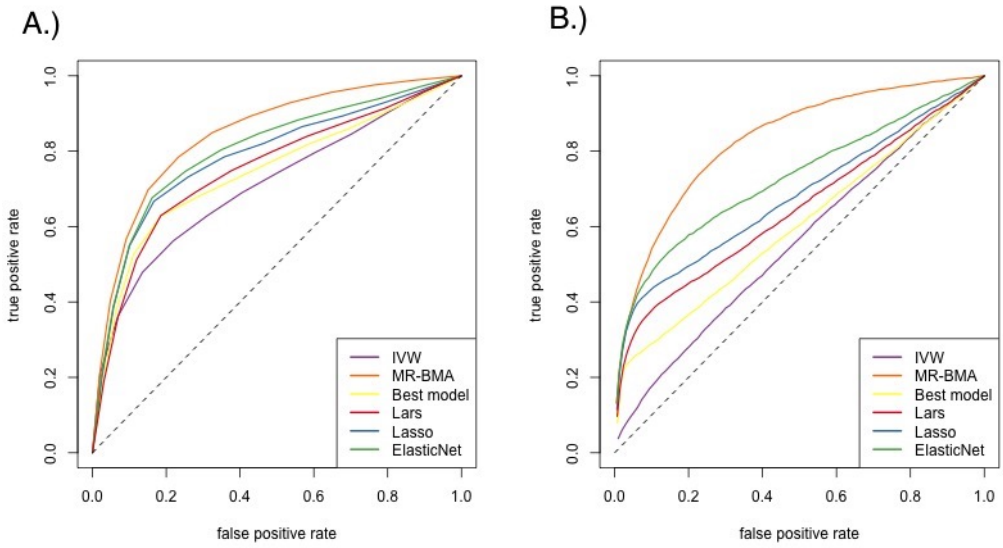

Figure 3: A) Receiver Operating Characteristic (ROC) curve for a small number of risk factors ($d = 12$) of which four have true positive effects (Scenario 1, Setting A). B) ROC curve for a large number of risk factors ($d = 92$) of which four have true positive effects (Scenario 2, Setting A). Proportion of variance explained (R^2) is set to 0.3.

	Scenario 1: NMR metabolites, $d = 12$ risk factors					
	Setting A (4 true risk factors, all risk increasing)			Setting B (8 true risk factors, risk increasing and decreasing)		
R^2	0.1	0.3	0.5	0.1	0.3	0.5
IVW	0.6727	0.1675	0.0784	0.5949	0.1619	0.0629
Lars	0.1292	0.0447	0.0298	0.1559	0.0648	0.0372
Lasso	0.0604	0.0289	0.0162	0.1046	0.0503	0.0307
Elastic Net	0.0673	0.0300	0.0162	0.1161	0.0480	**0.0287**
MR-BMA	**0.0340**	**0.0175**	**0.0105**	**0.0534**	**0.0368**	0.0306
Best model	0.0717	0.0320	0.0156	0.0921	0.0514	0.0376
	Scenario 2: NMR metabolites, $d = 92$ risk factors					
R^2	0.1	0.3	0.5	0.1	0.3	0.5
IVW	22.9516	6.0594	2.6257	23.2495	5.7715	2.4802
Lars	0.0354	0.0367	0.0094	0.0321	0.0212	0.0143
Lasso	0.0064	0.0047	0.0039	0.0105	0.0086	0.0074
Elastic Net	0.0064	0.0044	0.0034	0.0098	0.0078	0.0067
MR-BMA	**0.0051**	**0.0039**	**0.0032**	**0.0088**	**0.0076**	**0.0063**
Best model	0.0114	0.0081	0.0061	0.0150	0.0121	0.0096
	Scenario 3: blood cell traits, $d = 33$ risk factors					
R^2	0.1	0.3	0.5	0.1	0.3	0.5
IVW	1.6200	0.4272	0.1742	2.3377	0.6399	0.2770
Lars	0.3461	0.1151	0.0482	0.5971	0.1777	0.0960
Lasso	0.0161	0.0067	0.0040	0.0448	0.0315	0.0265
Elastic Net	0.0168	0.0074	0.0044	0.0526	0.0313	0.0270
MR-BMA	**0.0066**	**0.0034**	**0.0019**	**0.0307**	**0.0263**	**0.0239**
Best model	0.0128	0.0051	0.0027	0.0518	0.0348	0.0292

Table 1: Mean squared error (MSE) of the causal effect estimates from the competing methods on the NMR metabolite and blood cell trait data. We mark in bold the lowest MSE in each experimental setting.

Metabolites as risk factors for age-related macular degeneration

Next we demonstrate how MR-BMA can be used to select metabolites as causal risk factors for age-related macular degeneration (AMD). AMD is a painless eye-disease that ultimately leads to the loss of vision. AMD is highly heritable with an estimated heritability of up to 0.71 for advanced AMD in a twin study [20]. A GWAS meta-analysis has identified 52 independent common and rare variants associated with AMD risk at a level of genome-wide significance [9]. Several of these regions are linked to lipids or lipid-related biology, such as the *CETP*, *LIPC*, and *APOE* gene regions [21]. Lipid particles are deposited within drusen in the different layers of Bruch's membrane in AMD patients [21]. A recent observational study has highlighted strong associations between lipid metabolites and AMD risk [22].

This evidence for lipids as potential risk factor for AMD has motivated a multivariable MR analysis which has shown that HDL cholesterol may be a putative risk factor for AMD, while there was no evidence of a causal effect for LDL cholesterol and triglycerides [23]. Here, we extend this analysis to consider not just three lipid measurements, but a wider and more detailed range of $d = 30$ metabolite measurements to pinpoint potential causal effects more specifically. As summary-level data we use $d = 30$ metabolites as measured in the metabolite GWAS described earlier [11] for the same lipid-related instrumental variants as described previously. All of these metabolites have at least one genetic variant used as an instrumental variable that is genome-wide significant and no genetic associations of metabolites are stronger correlated

than $r = 0.985$. First, we prioritise and rank risk factors by their marginal inclusion probability (*MIP*) from MR-BMA using $\sigma^2 = 0.25$ as prior variance and $p = 0.1$ as prior probability, corresponding to *a priori* three expected causal risk factors. Secondly, we perform model diagnostics based on the best models with posterior probability > 0.02.

When including all genetic variants available in both the NMR and the AMD summary data ($n = 148$), the top risk factor with respect to its *MIP* (Supplementary Table 1 A) is LDL particle diameter (LDL.D, $MIP = 0.526$). All other risk factors have evidence less than $MIP < 0.25$. In order to check the model fit, we consider the best individual models (Supplementary Table 1 B) with posterior probability > 0.02. For illustration, we present here the predicted associations with AMD based on the best model including LDL.D, and TG content in small HDL (S.HDL.TG) against the observed associations with AMD. We colour code genetic variants according to their their q-statistic (Figure 4 A and Supplementary Figure 16 A, Supplementary Table 2) and Cook's distance (Figure 4 B and Supplementary Figure 16 B, Supplementary Table 3). First, the q-statistic indicates two variants, rs492602 in the *FUT2* gene region and rs6859 in the *APOE* gene region, as outliers in all best models. Second, the genetic variant with the largest Cook's distance ($Cd = 0.871$ to $Cd = 1.087$) consistently in all models investigated is rs261342 mapping to the *LIPC* gene region. This variant has been indicated previously to have inconsistent associations with AMD compared to other genetic variants [23, 24].

We repeat the analysis without the three influential and/or heterogeneous variants ($n = 145$), and report the ten risk factors with the largest marginal

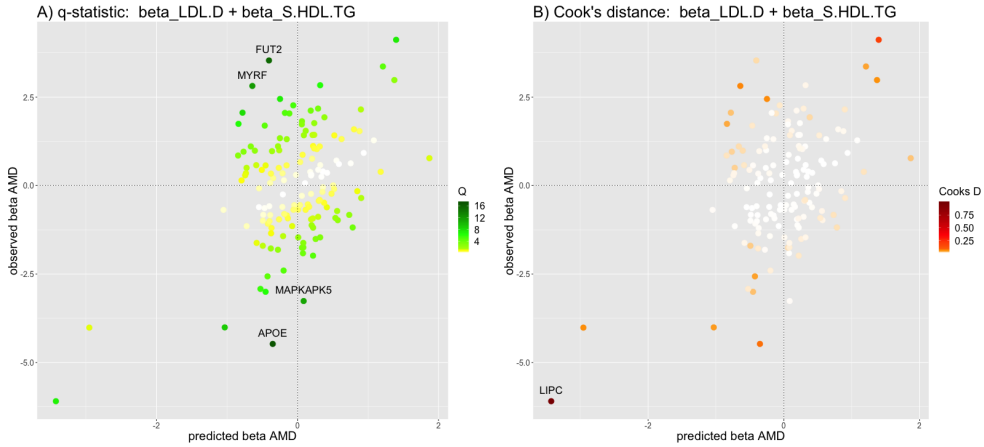

Figure 4: Diagnostic plot of the predicted associations with AMD based on the model including LDL.D, and S.HDL.TG (x-axis) against the observed associations with AMD (y-axis) showing all $n = 148$ genetic variants. This is the highest-ranking model when keeping outlying and influential genetic variants in the analysis. The colour code shows: left) the q-statistic for outliers and right) Cook's distance for the influential points. Any genetic variant with q-value larger than 10 or Cook's distance larger than the median of the relevant F-distribution is marked by a label indicating the gene region.

inclusion probability in Table 2 A and the full results in Supplementary Table 4. The top two risk factors are total cholesterol in extra-large HDL particles (XL.HDL.C, $MIP = 0.700$) and total cholesterol in large HDL particles (L.HDL.C, $MIP = 0.229$). XL.HDL.C and L.HDL.C were strongly correlated ($r = 0.80$), and models including both have very low evidence. Table 2 B gives the posterior probability of individual models. Supplementary Figure 17 shows the scatterplots of the genetic associations with each of these two risk factors individually against the genetic associations with AMD risk. We select the five individual models with a posterior probability > 0.02 to inspect the model fit (Supplementary Figures 18 and 19). This time, no genetic variant has a consistently large q-statistic (Supplementary Table 5) or

Cook's distance (Supplementary Table 6) . Repeating the analysis without the largest influential point, rs5880 in the *CETP* gene region, or the strongest outlier, rs103294 in the *AC245884.7* gene region, did not impact the ranking of the risk factors. We tested the robustness of the results with respect to a wide range of prior variance and prior probability parameters; results did not change substantially (Supplementary Tables 7 and 8).

We also applied Lars, Lasso, and Elastic Net after excluding outliers and influential points ($n = 145$). Lars showed the largest regression coefficient for L.HDL.C including eleven risk factors. Lasso selected four risk factors with the largest regression coefficient for XL.HDL.C, while Elastic Net selected ten risk factors with the largest regression coefficient for L.HDL.C. Full results for the competing methods are given in Supplementary Tables 9 to 12. A disadvantage of regularised regression approaches is that risk factor selection is binary; risk factors are either included in the model or set to have a coefficient zero. The magnitude of regularised regression coefficients does not rank risk factors according to their strength of evidence for inclusion in the model.

The detection of influential points in the initial analysis highlights rs26134, a genetic variant in the *LIPC* gene region, that had a strong impact on the analysis. Figure 5 shows the model diagnostics of the highest ranked model excluding outlying and influential points (XL.HDL.C as the sole risk factor), with the variant in the *LIPC* gene region also plotted. This particular variant exhibits a distinct, potentially pleiotropic, effect. While all other variants support that XL.HDL.C increases the risk of AMD, this particular variant has the opposite direction of association with AMD risk as that predicted by

its association with XL.HDL.C. Further functional and fine-mapping studies of this region are needed to understand the contrasting association of this variant with AMD risk.

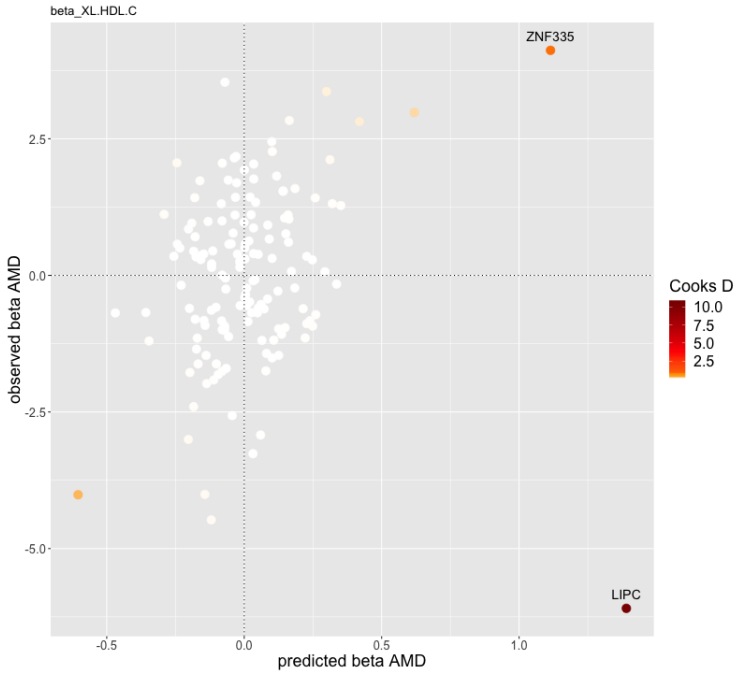

Figure 5: Diagnostic plot of the predicted associations with AMD based on the model including XL.HDL.C (x-axis) against the observed associations with AMD (y-axis) showing all $n = 148$ genetic variants, where the colour code shows Cook's distance for the genetic variants. This is the highest-ranking model on omission of outlying and influential genetic variants from the analysis. Note rs26134 in the *LIPC* gene region which has an anomalous direction of association with AMD risk in contrast to all other genetic variants.

These results confirm previous studies [23, 24] that identified HDL cholesterol as a putative risk factor for AMD and draw the attention to extra-large and large HDL particles. A recent observational study [22] supports our finding that extra-large HDL particles have an important role in the pathogenesis

of AMD.

As a further sensitivity analysis (detailed results not shown), we repeat this analysis with a different selection of instrumental variables using $n = 56$ independent genetic variants that were genome-wide hits for any metabolite measurement in this dataset [11]. Cholesterol content in large HDL particles is still selected with high posterior probability for this choice of variants underlining that the evidence for an effect of large HDL particles is not sensitive to the specific selection of lipid-related genetic variants.

Discussion

We here introduce a novel approach for multivariable MR, MR-BMA, which scales to the analysis of high-throughput experiments. This model averaging procedure prioritises and selects causal risk factors in a Bayesian framework from a high-dimensional set of related candidate risk factors. Our approach is especially suited for sparse settings, i.e. when the proportion of true causal risk factors compared to all risk factors considered is small. As is common for statistical techniques for variable selection, MR-BMA does not provide unbiased estimates. However, as shown in the simulation study, causal estimates from MR-BMA have reduced variance and thus MR-BMA improves over unbiased approaches, like the IVW method, in terms of mean squared error and detection of true risk factors. The primary aim of this work is to detect causal risk factors rather than to unbiasedly estimate the magnitude of their causal effects.

We demonstrated the approach with application to a dataset of NMR

metabolites, which included predominantly lipid measurements, using variants associated with lipids as instrumental variables. Previous MR analysis [23, 24] including three lipid measurements from the Global Lipids Genetics Consortium [19] have identified HDL cholesterol as potential risk factor for AMD. Our new approach to multivariable MR refined this analysis and confirmed HDL cholesterol as a potential causal risk factor for AMD, further pinpointing that large or extra-large HDL particles are likely to be driving disease risk. Other areas of application where this method could be used include imaging measurements of the heart and coronary artery disease, body composition measures and type 2 diabetes, or blood cell traits and atherosclerosis. As multivariable MR accounts for measured pleiotropy, this approach facilitates the selection of suitable genetic variants for causal analyses. In each case, it is likely that genetic predictors of the set of risk factors can be found, even though finding specific predictors of, for example, particular heart measurements from cardiac imaging, may be difficult given widespread pleiotropy [25]. This approach allows a more agnostic and hypothesis-free approach to causal inference, allowing the data to identify the causal risk factors.

Multivariable MR estimates the direct effect of a risk factor on the outcome and not the total effect as estimated in standard univariable MR. This is in analogy with multivariable regression where the regression coefficients represent the association of each variable with the outcome when all others are held constant. Having said this, the main goal of our approach is risk factor selection, and not the precise estimation of causal effects, since the variable selection procedure shrinks estimates towards the null. If there are

mediating effects between the risk factors, then this approach will identify the risk factor most proximal to and has the most direct effect on an outcome. For example, if the risk factors included would form a signalling cascade then our approach would identify the downstream risk factor in the cascade with the direct effect on the outcome and not the upstream risk factors in the beginning of the cascade. Hence, a risk factor may be a cause of the outcome, but if its causal effect is mediated via another risk factor included in the analysis, then it will not be selected in the multivariable MR approach.

Our approach is formulated in a Bayesian framework. Particular care needs to be taken when choosing the hyper-parameter for the prior probability which relates to the *a priori* expected number of causal risk factors. In the applied example the results were robust to a wide range of prior specifications for the parameter as seen in Supplementary Table 7. Additionally, the prior variance of the causal parameters needs to be specified and tested for robustness as we show in the Supplementary Table 8.

When genetic variants are weak predictors for the risk factors, this can introduce weak instrument bias. In univariable two-sample MR, any bias due to weak instruments is towards the null and does not lead to inflated type 1 error rates [26]. However, in multivariable MR, weak instrument bias can be in any direction (Methods), although bias will tend to zero as the sample size increases. Selection of risk factors is only possible if there are genetic variants that are predictors of these risk factors. Consequently, we need to be cautious about the interpretation of null findings, particularly in our example for non-lipid risk factors, as these might be deprioritised in terms of statistical power by our choice of genetic variants. One of the

biggest challenges of multivariable MR is the design of a meaningful study, in particular the choice of both, the genetic variants and the risk factors. The design of the study is important for the interpretation of the risk factors prioritised: The ranking of risk factors is conditional on the genetic variants used. For instance, in our applied example we find evidence for extra large and large HDL cholesterol concentration given that we used lipid-related genetic variants as instrumental variables. We recommend to include only risk factors which have at least one, and ideally multiple genetic variants that act as strong instruments. Caution is needed for the interpretation of null findings, particularly in our example for non-lipid risk factors, as these might be deprioritised in terms of statistical power by our choice of genetic variants.

A further requirement for multivariable MR is that the genetic variants can distinguish between risk factors [13]. We recommend to check the correlation structure between genetic associations for the selected genetic variants and to include no pair of risk factors which is extremely strongly correlated. In the applied example, we included only risk factors with an absolute correlation less than 0.99. As we were not able to include more than three measurements for each lipoprotein category (cholesterol content, triglyceride content, diameter), care should be taken not to overinterpret findings in terms of the specific measurements included in the analysis rather than those correlated measures that were excluded from the analysis (such as phospholipid and cholesterol ester content).

Another assumption of multivariable MR is that there is no unmeasured horizontal pleiotropy. This means that the variants do not influence the

outcome except via the measured risk factors. The assumption of no horizontal pleiotropy is a common and untestable assumption in MR. It is an active area of research to robustify MR against violations of this assumption. Some of these robust methods for MR make a specific assumption about the behaviour of pleiotropic variants, such as MR-Egger [27], which assumes pleiotropic effects are uncorrelated from the genetic associations with the risk factor – the InSIDE assumption. Other methods exclude outlying variants as they are potentially pleiotropic such as MR-PRESSO [28]. In multivariable MR, pleiotropic variants can be detected as outliers to the model fit. Here we quantify outliers using the q-statistic. Outlier detection in standard univariable MR can be performed by model averaging where different subsets of instruments are considered [29, 30], assuming that a majority of instruments is valid, but without prior knowledge which are the valid instruments. In multivariable MR, ideally one would like to perform model selection and outlier detection simultaneously. Additionally, we search for genetic variants that are influential points. While these may not necessary be pleiotropic, we suggest removing such variants as a sensitivity analysis to judge whether the overall findings from the approach are dominated by a single variant. Findings are likely to be more reliable when they are evidenced by multiple genetic variants.

In conclusion, we introduce here MR-BMA, the first approach to perform risk factor selection in multivariable MR, which can identify causal risk factors from a high-throughput experiment. MR-BMA can be used to determine which out of a set of related risk factors with common genetic predictors are the causal drivers of disease risk.

Methods

Methods is available online. The Supplementary Information includes Supplementary Note S1 that describes the derivation of the Bayes Factors and one Supplementary Material providing Supplementary Tables and Figures to support the simulation study and application.

Web resources

MR-BMA and publicly available summary data on AMD and NMR metabolites as presented in the applied example is public on `https://github.com/verena-zuber/demo_AMD`. We provide R-code and documentation to reproduce the results and figures of the applied example.

A) Model averaging

	Risk factor	Marginal inclusion probability (MIP)	Model-averaged causal estimate $\hat{\theta}_{\text{MACE}}$
1	XL.HDL.C	0.700	0.344
2	L.HDL.C	0.229	0.087
3	HDL.D	0.087	0.022
4	XS.VLDL.TG	0.082	-0.019
5	LDL.D	0.074	-0.018
6	IDL.TG	0.066	-0.012
7	XXL.VLDL.TG	0.063	0.018
8	S.VLDL.TG	0.062	-0.014
9	Serum.TG	0.061	-0.014
10	Serum.C	0.054	-0.011

B) Individual models

	Risk factor(s)	Posterior probability (PP)	Model-specific causal estimates $\hat{\theta}_{\gamma}$
1	XL.HDL.C	0.156	0.509
2	L.HDL.C	0.078	0.384
3	XL.HDL.C,XS.VLDL.TG	0.026	0.457,-0.181
4	IDL.TG,XL.HDL.C	0.025	-0.179,0.495
5	HDL.D	0.023	0.359
6	Serum.C,XL.HDL.C	0.019	-0.183,0.573
7	S.VLDL.TG,XL.HDL.C	0.015	-0.172,0.443
8	S.VLDL.C,XL.HDL.C	0.014	-0.164,0.477
9	Serum.TG,XL.HDL.C	0.014	-0.169,0.465
10	S.HDL.TG,XL.HDL.C	0.013	-0.18,0.415

Table 2: Ranking of risk factors for age-related macular degeneration (AMD) after exclusion of outlying and influential variants ($n = 145$): A) according to their marginal inclusion probability (MIP) and B) the best ten individual models according to their posterior probability (PP). Results are given after excluding the *APOE*, *FUTC*, and *LIPC* regions. $\hat{\theta}_{\text{MACE}}$ is the model averaged causal effect of a risk factor and $\hat{\theta}_{\gamma}$ is the causal effect estimate for a specific model. Abbreviations: HDL.D = HDL diameter, IDL.TG = Triglycerides in IDL, L.HDL.C = Total cholesterol in large HDL, LDL.D = LDL diameter, Serum.C = Serum total cholesterol, Serum.TG = Serum total triglycerides, S.VLDL.C =Total cholesterol in small VLDL, S.VLDL.TG = Triglycerides in small VLDL, XS.VLDL.TG = Triglycerides in very small VLDL, XL.HDL.C = Total cholesterol in very large HDL

Acknowledgements

This work was supported by the UK Medical Research Council (MC_UU_00002/7). S.B. and V.Z. are supported by Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 204623/Z/16/Z). This study would not have been possible without the access to publicly available summary data. We would like to thank the International AMD Genetics consortium (`http://amdgenetics.org/`), the authors of the blood trait GWAS as curated by the GWAS catalog (`https://www.ebi.ac.uk/gwas/`), and the authors of the NMR GWAS (`http://www.computationalmedicine.fi/data`).

References

[1] Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1–22 (2003). `http://ije.oxfordjournals.org/cgi/reprint/32/1/1.pdf`.

[2] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian Randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163 (2008). URL `https://doi.org/10.1002/sim.3034`.

[3] Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018). URL `https://doi.org/10.1038/s41586-018-0175-2`.

[4] Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/27863252`.

[5] Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nature Genetics* **46**, 543 EP – (2014). URL http://dx.doi.org/10.1038/ng.2982.

[6] Biffi, C. *et al.* Three-dimensional cardiovascular imaging-genetics: a mass univariate framework. *Bioinformatics* **34**, 97–103 (2018). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870605/.

[7] the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121 EP – (2015). URL http://dx.doi.org/10.1038/ng.3396.

[8] Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature Genetics* **50**, 559–571 (2018). URL https://doi.org/10.1038/s41588-018-0084-1.

[9] Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics* **48**, 134 EP – (2015). URL http://dx.doi.org/10.1038/ng.3448.

[10] Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology* **181**, 251–260 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25632051.

[11] Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications* **7** (2016). URL http://dx.doi.org/10.1038/ncomms11122.

[12] Burgess, S. *et al.* Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. *Genetics* **207**, 481–487 (2017). URL http://www.genetics.org/content/207/2/481.

[13] Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian Randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology* dyy262–dyy262 (2018). URL http://dx.doi.org/10.1093/ije/dyy262.

[14] Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med* (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28960498.

[15] Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557–560 (2003). URL https://www.bmj.com/content/327/7414/557.

[16] Cook, R. D. Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174 (1979). URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481634.

[17] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Statist.* **32**, 407–499 (2004). URL https://projecteuclid.org:443/euclid.aos/1083178935.

[18] Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**, 1–22 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20808728.

[19] Consortium, G. L. G. Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274 EP – (2013). URL http://dx.doi.org/10.1038/ng.2797.

[20] Seddon, J. M., Cote, J., Page, W. F., Aggen, S. H. & Neale, M. C. The US Twin Study of Age-Related Macular Degeneration: Relative Roles of Genetic and Environmental Influences. *Archives of Ophthalmology* **123**, 321–327 (2005). URL https://doi.org/10.1001/archopht.123.3.321.

[21] van Leeuwen, E. M. *et al.* A new perspective on lipid research in age-related macular degeneration. *Progress in Retinal and Eye Research* **67**, 56 – 86 (2018). URL http://www.sciencedirect.com/science/article/pii/S1350946217301271.

[22] Colijn, J. *et al.* Increased High Density Lipoprotein-levels associated with Age-related Macular degeneration. Evidence from the EYE-RISK and E3 Consortia. *Ophthalmology* (2018). URL http://www.sciencedirect.com/science/article/pii/S0161642018310911.

[23] Burgess, S. & Davey Smith, G. Mendelian Randomization Implicates High-Density Lipoprotein Cholesterol–Associated Mechanisms in Etiology of Age-Related Macular Degeneration. *Ophthalmology* (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28456421.

[24] Fan, Q. *et al.* HDL-cholesterol levels and risk of age-related macular degeneration: a multiethnic genetic study using Mendelian randomization. *International Journal of Epidemiology* **46**, 1891–1902 (2017). URL http://dx.doi.org/10.1093/ije/dyx189.

[25] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483 EP – (2013). URL http://dx.doi.org/10.1038/nrg3461.

[26] Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology* **178**, 1177–1184 (2013). URL http://dx.doi.org/10.1093/aje/kwt084.

[27] Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015). URL http://dx.doi.org/10.1093/ije/dyv080.

[28] Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian Randomization between complex traits and diseases. *Nature Genetics* **50**, 693–698 (2018). URL https://doi.org/10.1038/s41588-018-0099-7.

[29] Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian Randomization via the zero modal pleiotropy assumption. *International journal of epidemiology* **46**, 1985–1998 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29040600.

[30] Burgess, S., Zuber, V., Gkatzionis, A. & Foley, C. N. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian Randomization when a plurality of candidate instruments are valid. *International Journal of Epidemiology* dyy080–dyy080 (2018). URL http://dx.doi.org/10.1093/ije/dyy080.

Methods

Mendelian Randomization data input: Summarized data set-up

One of the key features of Mendelian Randomization (MR) is that the approach can be performed using summarised data on genetic associations – beta-coefficients and their standard errors from univariate regression analyses. No access to individual-level genotype data is needed. Additionally, these association estimates can be derived from different samples. In two-sample MR, the genetic associations with the risk factor are derived from one sample and the genetic associations with the outcome from another sample [1]. The use of summarised data in two-sample MR allows the sample size to be maximised by integrating data from large meta-analyses including hundreds of thousands of participants.

We assume the context of two-sample MR with summarized data [2]. For each genetic variant $i = 1, \ldots, n$ and each risk factor $j = 1, \ldots, d$, we take the beta-coefficient $\beta^{\star}_{X_{ij}}$ and standard error $\text{se}(\beta^{\star}_{X_{ij}})$ from a univariable regression in which the risk factor X_j is regressed on the genetic variant G_i in sample one, and beta-coefficient $\beta^{\star}_{Y_i}$ and standard error $\text{se}(\beta^{\star}_{Y_i})$ from a univariable regression in which the outcome Y is regressed on the genetic variant G_i in sample two. For simplicity of notation, although the beta-coefficients are estimates, we omit the conventional "hat" notation and treat the beta-coefficients as observed data points. When considering multiple risk factors, we construct a matrix of beta-coefficients $\beta^{\star}_{X}$ of dimension $n \times d$, where d is

the number of risk factors and n is the number of genetic variants.

We assume that the genetic effects on risk factors and on the outcome are linear and homogeneous across the population, and identical between the two samples [3]. Furthermore, we assume that the n genetic variants selected as instrumental variables are independent, an assumption common in MR studies. This is usually achieved by including only the lead genetic variant from each gene region in the analysis. Finally, we assume that genetic association estimates are derived from two distinct samples with no overlap between the samples. These assumptions can all be relaxed to some extent if the goal is causal inference rather than causal estimation; see [4] for details.

Multivariable Mendelian randomization and the linear model

Multivariable MR is an extension of the standard MR paradigm (Figure 1) to model not one, but multiple risk factors as illustrated in Figure 2. Univariable MR can be cast as a weighted linear regression model in which the genetic associations with the outcome $\beta^\star_{Y_i}$ are regressed on the genetic associations with the risk factor $\beta^\star_{X_i}$ [5]

$$\beta^\star_{Y_i} = \theta\beta^\star_{X_i} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \mathrm{se}(\beta^\star_{Y_i})^2). \tag{1}$$

In multivariable MR, the genetic associations with the outcome are regressed on the genetic associations with all the risk factors [6]

$$\beta^\star_{Y_i} = \theta_1\beta^\star_{X_{i1}} + \theta_2\beta^\star_{X_{i2}} + \ldots + \theta_d\beta^\star_{X_{id}} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \mathrm{se}(\beta^\star_{Y_i})^2). \tag{2}$$

Weights in these regression models are proportional to inverse of the variance of the genetic association with the outcome ($\text{se}(\beta^{\star}_{Y_i})^{-2}$). This is to ensure that genetic variants having more precise association estimates receive more weight in the analysis. The same weighting can also be achieved by standardising the association estimates, by dividing $\beta^{\star}_{Y_i}$ and $\beta^{\star}_{X_i}$ by $\text{se}(\beta^{\star}_{Y_i})$. In the following derivations, we assume that $\beta_Y = \beta^{\star}_{Y_i}/\,\text{se}(\beta^{\star}_{Y_i})$ and $\beta_{X_i} = \beta^{\star}_{X_i}/\,\text{se}(\beta^{\star}_{Y_i})$ are standardised, so that the variances of the ϵ_i terms are all 1. To account for heterogeneity in the regression equation, we can use a multiplicative random effects model, which increases the variance of the error terms by a multiplicative factor [7]. Our parameter of interest is the vector of regression coefficients $\boldsymbol{\theta} = \{\theta_1, ..., \theta_d\}$. These are the direct causal effects of the risk factors in turn on the outcome when all the other risk factors in the model are held constant [8]. In contrast, univariable Mendelian randomization using genetic variants that are instrumental variables for the specific risk factor of interest estimates the total effect of the risk factor on the outcome. The direct effect will differ from the total effect if the effect of the risk factor is mediated via another risk factor included in the model [9]. We illustrate the difference between the direct and total effect using directed acyclic graphs in Supplementary Figure 1. In some cases (such as to identify the proximal risk factor to the outcome), the direct effect is of interest; in other cases (such as to evaluate the potential impact of intervening on a risk factor), it is the total effect that is truly of interest [8].

Choosing genetic variants as instruments

In multivariable MR, a genetic variant is a valid instrumental variable if the following criteria hold:

- IV1 Relevance: The variant is associated with at least one of the risk factors.
- IV2 Exchangeability: The variant is independent of all confounders of each of the risk factor–outcome associations.
- IV3 Exclusion restriction: The variant is independent of the outcome conditional on the risk factors and confounders.

One of the main differences of multivariable MR compared to univariable MR is the relaxation of the exclusion restriction condition. In contrast to univariable MR, multivariable MR allows for measured pleiotropy [10] via any of the observed risk factors. Hence the instrumental variable assumptions are more likely to be satisfied for multivariable MR than for univariable MR for a given choice of genetic variants.

It is not necessary for every genetic variant to be associated with all the risk factors, although if no genetic variants are associated with a particular risk factor, then the causal effect of that risk factor cannot be identified. This would also occur if the genetic associations with two risk factors were exactly proportional. For precise identification of causal risk factors, it is necessary to have some variants that are more strongly associated with particular risk factors than others [9]. More precisely a risk factor can be included into the analysis if the following criteria (RF1-RF2) hold:

- RF1 Relevance: The risk factor needs to be strongly instrumented by at least one genetic variant included as instrumental variable.

- RF2 No multi-collinearity: The genetic associations of any risk factor included cannot be linearly explained by the genetic associations of any other risk factor or by the combination of genetic associations of multiple other risk factors included in the analysis.

We initially assume that all genetic variants are valid instruments. There is an emerging literature [11, 12] on how to perform robust MR analysis in the presence of invalid instruments; similar extensions can be adapted for multivariable MR [10].

Risk factor selection as variable selection in the linear model

We consider the situation in which we have a set of genetic variants that are instrumental variables for a set of risk factors, and we want to select which of those risk factors are causes of the outcome. Our implicit prior belief is that not all of the risk factors are causally related to the outcome and that the set of true causal risk factors is sparse. We formulate the selection of risk factors in two-sample multivariable MR as a variable selection task in the linear regression framework. In order to model the correlation between risk factors we base our likelihood on a Gaussian distribution

$$\beta_Y \mid \boldsymbol{\beta}_{\mathbf{X}}, \boldsymbol{\theta}, \tau \sim N(\boldsymbol{\beta}_{\mathbf{X}}\boldsymbol{\theta}, \frac{1}{\tau}). \tag{3}$$

Following the D_2 prior specifications as introduced in [13], we use the following conjugate priors for the causal effects $\boldsymbol{\theta}$, the residual error ϵ, and the precision τ

$$\begin{aligned} \boldsymbol{\theta} &\sim N(0, \boldsymbol{\nu}/\tau) \\ \epsilon &\sim N(0, \frac{1}{\tau}) \\ \tau &\sim \Gamma(\kappa/2, \lambda/2), \end{aligned} \tag{4}$$

where $\boldsymbol{\nu} = diag(\sigma^2)$ is the diagonal variance matrix of the causal effects (independence prior), and the precision τ is assumed to follow a Gamma distribution with hyperparameters κ as the shape and λ as the scale parameter. Next, we introduce a binary indicator $\boldsymbol{\gamma}$ of length d that indicates which risk factors are selected and which ones are not

$$\gamma_j = \begin{cases} 1, \text{ if the } j\text{th risk factor is selected,} \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

The indicator $\boldsymbol{\gamma}$ encodes a specific regression model $M_{\boldsymbol{\gamma}}$ that includes the risk factors as indicated in $\boldsymbol{\gamma}$. A model $M_{\boldsymbol{\gamma}}$ can include one or a combination of multiple risk factors. To evaluate the evidence of a specific model $M_{\boldsymbol{\gamma}}$, we calculate the Bayes factor for model $M_{\boldsymbol{\gamma}}$ against the null model that does not include an intercept or any risk factor. The Bayes factor $BF(M_{\boldsymbol{\gamma}})$ has the following closed form representation

$$BF(M_{\boldsymbol{\gamma}}) = \frac{|\boldsymbol{\Omega}|^{1/2}}{|\boldsymbol{\nu}_{\boldsymbol{\gamma}}|^{1/2}} \left(\frac{\beta_Y^t \beta_Y - \boldsymbol{\Theta}^t \boldsymbol{\Omega}^{-1} \boldsymbol{\Theta}}{\beta_Y^t \beta_Y} \right)^{-n/2}, \tag{6}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Omega}\boldsymbol{\beta}_{\mathbf{X}_\gamma}{}^t\beta_Y$ is the causal effect estimate and $\boldsymbol{\Omega} = (\boldsymbol{\nu}_{\boldsymbol{\gamma}}^{-1} + \boldsymbol{\beta}_{\mathbf{X}_\gamma}{}^t\boldsymbol{\beta}_{\mathbf{X}_\gamma})^{-1}$ is the inverse of the shrinkage covariance between the genetic associations of the risk factors. For a detailed derivation of the Bayes factor we refer to the Supplementary Note S1.

Prior specification

Another important aspect is the prior for the model size k, which we model using a Binomial distribution

$$Pr(K = k) = \binom{d}{k} p^k (1-p)^{d-k}. \tag{7}$$

This requires choosing the probability p of including a risk factor in the model according to prior assumptions regarding the sparsity of the results. We recommend to select p according to the expected *a priori* model size, which is $p \times d$. Currently, all risk factors are assumed to have the same prior probability, and thus the probability of all models of the same size k is equal. The prior of a specific model $M_{\boldsymbol{\gamma}}$ of size k is defined as

$$p(M_{\boldsymbol{\gamma}}) = \binom{d}{k}^{-1} Pr(K = k) = p^k (1-p)^{d-k}. \tag{8}$$

The second important aspect is the prior for the variance of the risk factors $\boldsymbol{\nu} = diag(\sigma^2)$, where we assume that all risk factors have the same prior variance σ^2. Large values of σ^2 would favour strong causal effects of the risk factors on the outcome. Following [13] we initially set $\sigma^2 = 0.25$, but sensitivity of the results with respect to this prior should be investigated. The

parameter can be specified in the implementation of MR-BMA. In the applied example we perform a sensitivity analysis for this important parameter.

Posterior calculation and marginal inclusion probability of a risk factor

Let $\mathbf{\Gamma}$ be the space of all possible combinations of risk factors. The posterior probability (PP) of a model $M_{\boldsymbol{\gamma}}$ can be expressed by the prior probability (8) and the Bayes factor (6) of model $M_{\boldsymbol{\gamma}}$ is

$$PP(M_{\boldsymbol{\gamma}} \mid \beta_Y, \boldsymbol{\beta}_{\mathbf{X}}) = \frac{p(M_{\boldsymbol{\gamma}})BF(M_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}\in\mathbf{\Gamma}} p(M_{\boldsymbol{\gamma}})BF(M_{\boldsymbol{\gamma}})}. \tag{9}$$

In high-dimensional variable selection, the evidence for one particular model can be small because the model space is very large and many models might have comparable evidence. This is why MR-BMA uses Bayesian model averaging (BMA) and computes for each risk factor j its marginal inclusion probability (MIP), which is defined as the sum of the posterior probabilities over all models where the risk factor is present

$$MIP(j = 1 \mid \beta_Y, \boldsymbol{\beta}_{\mathbf{X}}) = \frac{\sum_{\boldsymbol{\gamma}\in\mathbf{\Gamma}} I(\gamma_j = 1)p(M_{\boldsymbol{\gamma}})BF(M_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}\in\mathbf{\Gamma}} p(M_{\boldsymbol{\gamma}})BF(M_{\boldsymbol{\gamma}})}, \tag{10}$$

where $I(\gamma_j = 1)$ equals 1 if risk factor j is part of the model and 0 otherwise.

An exhaustive evaluation of all possible combinations of risk factors is computationally prohibitive already for a moderate number of risk factors ($d > 20$). To alleviate this issue we have implemented a shotgun stochastic search algorithm [14] that evaluates all combinations of risk factors with a

non-negligible contribution to the calibration factor $\sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} p(M_{\boldsymbol{\gamma}})BF(M_{\boldsymbol{\gamma}})$ in equation (9). This algorithm is based on the assumption that the majority of combinations of risk factors have a posterior probability close to zero and do not need to be considered when computing the calibration factor in the denominator of equations (9) and (10).

Causal estimation

We derive the estimates for the causal effects $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ of model $M_{\boldsymbol{\gamma}}$ as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} = \boldsymbol{\Omega}\boldsymbol{\beta}_{\mathbf{X}_{\boldsymbol{\gamma}}}{}^{t}\beta_Y = (\boldsymbol{\nu}_{\boldsymbol{\gamma}}^{-1} + \boldsymbol{\beta}_{\mathbf{X}_{\boldsymbol{\gamma}}}{}^{t}\boldsymbol{\beta}_{\mathbf{X}_{\boldsymbol{\gamma}}})^{-1}\boldsymbol{\beta}_{\mathbf{X}_{\boldsymbol{\gamma}}}{}^{t}\beta_Y, \tag{11}$$

which is closely related to the regression coefficient in Ridge regression. Adding the diagonal matrix $\boldsymbol{\nu}_{\boldsymbol{\gamma}}^{-1}$ stabilises the inversion and makes the estimate more robust to strong correlation among risk factors. There can be strong correlation between candidate risk factors as seen in the genetic correlation matrices in the applied examples as illustrated in Supplementary Figure 2 and 11, which makes it important to stabilise the causal estimate.

The model-averaged causal estimate (MACE) for risk factor j from the MR-BMA approach is

$$\hat{\boldsymbol{\theta}}_{\text{MACE}}(j) = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} I(\gamma_j = 1)PP(M_{\boldsymbol{\gamma}} \mid \beta_Y, \boldsymbol{\beta}_{\mathbf{X}})\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}. \tag{12}$$

MR-BMA ranks and prioritises risk factors according to their marginal inclusion probability and estimates the MACE as defined in equation (12). As an alternative approach, we also consider selecting the 'best model' based

on the individual model posterior probabilities as defined in equation (9).

Detection of invalid and influential instruments

Invalid instruments may be detected as outliers with respect to the fit of a specific linear model $M_{\boldsymbol{\gamma}}$. We recommend to check the best individual models for outliers by visual inspection of the scatterplot of the predicted associations based on $M_{\boldsymbol{\gamma}}$ with the outcome $\hat{\boldsymbol{\beta}}_Y = \boldsymbol{\beta}_{\mathbf{X}_{\boldsymbol{\gamma}}}\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ against the actual observed β_Y. If a genetic variant is detected consistently as an outlier in several of the top models, it may be advisable to explore the analyses excluding that outlying variant from the analysis. To quantify outliers we use the Q-statistic, which is an established tool for identifying heterogeneity in meta-analysis [15]. It is defined as the sum of the residual vector q, which is the squared difference between the observed and predicted association with the outcome

$$\mathbf{Q} = \sum_i q_i = \sum_i (\beta_{Y_i} - \hat{\beta}_{Y_i})^2. \tag{13}$$

We note that equation 13 is defined on the weighted coefficients β_{Y_i}. When considering the unweighted coefficients $\beta^{\star}_{Y_i}$ the Q-statistic [9] is defined as

$$\mathbf{Q} = \sum_i q_i = \sum_i \frac{1}{\mathrm{se}(\beta^{\star}_{Y_i})^2}(\beta^{\star}_{Y_i} - \hat{\beta}^{\star}_{Y_i})^2, \tag{14}$$

with first order weighting equal to $\frac{1}{\mathrm{se}(\beta^{\star}_{Y_i})^2}$ [16].

The individual element q_i measures the heterogeneity of genetic variant i for a particular model $M_{\boldsymbol{\gamma}}$. We refer to q_i as the q-statistic, and use this to evaluate if specific genetic variants are outliers to the model fit.

Even if there are no outliers, it is advisable to check for influential observations and re-run the approach omitting a particular influential variant from the analysis. If a particular genetic variant has a strong association with the outcome, then it may have undue influence on the variable selection, leading to a model that fits that particular observation well, but other observations poorly. To quantify influential observations for a particular model $M_{\boldsymbol{\gamma}}$ we suggest to use Cook's distance [17]

$$Cd_i = \frac{q_i}{s^2 d} \frac{h_i}{(1-h_i)^2}, \tag{15}$$

where h_i is the ith diagonal element of the hat matrix $\mathbf{H} = \boldsymbol{\beta}_{\mathbf{X}_\gamma}(\boldsymbol{\nu}_{\boldsymbol{\gamma}}^{-1} + {\boldsymbol{\beta}_{\mathbf{X}_\gamma}}^t \boldsymbol{\beta}_{\mathbf{X}_\gamma})^{-1} {\boldsymbol{\beta}_{\mathbf{X}_\gamma}}^t$, and $s^2 = \frac{1}{n-d}\epsilon^t\epsilon$ is the mean squared error of the regression model. Following [18], we recommend to use the median of a central F-distribution with d and $n-d$ degrees of freedom as a threshold, and remove variants that have a Cook's distance which exceeds this value.

Impact of weak instrument bias

In the following presentation, we consider two risk factors with observed genetic associations β_{X_1} and β_{X_2}, which are a sum of the true genetic associations $\beta^\dagger_{X_1}$ and $\beta^\dagger_{X_2}$ and an additional error term ϵ_1 and ϵ_2 respectively, i.e.

$$\begin{aligned} \beta_{X_1} &= \beta^\dagger_{X_1} + \epsilon_1 \\ \beta_{X_2} &= \beta^\dagger_{X_2} + \epsilon_2. \end{aligned}$$

From this we define $\lambda_1 = \frac{var(\epsilon_1)}{var(\beta^{\dagger}_{X_1})}$ as the ratio of the uncertainty in the estimates of the genetic associations ($var(\epsilon_1)$) over the variability of the true genetic associations $var(\beta^{\dagger}_{X_1})$, and we define λ_2 similarly. Further let ρ be the correlation between β_{X_1} and β_{X_2}, and let θ_1 and θ_2 be the true direct effect of X_1 on Y and X_2 on Y, respectively. Following the measurement error literature [19], we derive the induced bias of the IVW estimates of the true causal effects θ_1 and θ_2, respectively, as

$$\begin{aligned}
\hat{\theta}_1 &= \theta_1 - \frac{\theta_1\lambda_1 - \rho\theta_2\lambda_2}{1-\rho} \\
\hat{\theta}_2 &= \theta_2 - \frac{\theta_2\lambda_2 - \rho\theta_1\lambda_1}{1-\rho},
\end{aligned}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the expected values of the effects for the mismeasured genetic association estimates. Looking closer at λ, the variability across variants of the true genetic associations, $var(\beta^{\dagger}_{X})$, is related to instrument strength. Thus the induced bias will be smaller the stronger the instruments. At the same time the uncertainty of the genetic association estimates, $var(\epsilon)$, decreases when increasing the sample size. If the genetic associations with the risk factors are estimated with different degrees of uncertainty, then bias could be more considerable. Analogous to differential measurement error, risk factors with more precisely estimated genetic associations would be prioritized in the regression model. In our application, all risk factors are measured on the same high-throughput platform and on the same sample size, thus reducing the impact of weak instrument bias to influence the ranking of risk factors.

Simulation study

To evaluate the performance of MR-BMA, we perform a simulation study taking genetic associations with risk factors from two real datasets, the first one based on genetic associations with NMR metabolites [20] and secondly on genetic associations with blood cell traits [21]. Further information on the data sets and pre-processing is given in the next sections. We simulate genetic associations with the outcome β_Y based on a subset of risk factors selected at random, which we refer to as the 'true' risk factors. We investigate three different scenarios and six sets of parameter values per scenario:

- Size of the data set: small ($d = 12$ NMR metabolites selected at random), large ($d = 92$ all NMR metabolites available), and moderate ($d = 33$ all blood cell traits available) number of risk factors included.
- Number of true risk factors: Setting A) four risk factors have an effect of $\theta = 0.3$, the other risk factors have no effect. Setting B) four risk factors have an effect of $\theta = 0.3$, and another four risk factors have an effect of $\theta = -0.3$, the other risk factors have no effect.
- Proportion of variance in the outcome explained by the risk factors: $R^2 = 0.1, 0.3, 0.5$ which defines the variance of the error.

We compare six different analysis methods:

- Multivariable inverse variance weighted (IVW) regression (equation 2) [6]
- Least-angle regression (Lars) as L1 regularised regression [22]

- Lasso as L1 regularised regression [23]
- Elastic Net as L1 and L2 regularised regression [23]
- MR-BMA using marginal inclusion probabilities
- Bayesian best model selection using posterior probabilities of individual models

Both Lars [22] and Lasso are versions of L1 regularised linear regression, and Elastic Net is a mixture of a L1 and L2 regularised linear regression, all of which have been devised for variable selection in high-dimensional data. We use here the Lars implementation [22] and for Lasso and Elastic Net we use the glmnet [23] implementation. For all regularised regression methods, we use cross-validation (CV) to tune the regularisation parameter to achieve the minimum cross-validation MSE. For the small risk factor space including 12 NMR metabolites, the MR-BMA approach is performed using an exhaustive search of all possible models with prior probability of a risk factor to be included set to $p = 0.5$, while for the moderate and large risk factor space of $d = 33$ blood cell traits and $d = 92$ NMR metabolites we employ the stochastic search with 10,000 iterations and $p = 0.1$. This reflects an expected *a priori* model size of six for the small risk factor space and around three for the blood cell traits and nine for the high-dimensional NMR metabolite setting. The prior variance σ^2 is fixed to 0.25.

Data pre-processing for NMR metabolites for simulation

The first data resource used for the simulation and application is publicly-available summarized data on genetic associations with risk factors derived

from a NMR metabolite GWAS [20] from `http://computationalmedicine.fi/data#NMR_GWAS`. All of the metabolites were inverse rank-based normal transformed, so the association estimates are all in standard deviation units. In order to avoid selection bias, we choose genetic variants based on an external data-set. As the majority of the metabolite measures relates to lipids, we take $n = 150$ independent genetic variants that are associated with any of three composite lipid measurements (LDL-cholesterol, triglycerides, or HDL-cholesterol) at a genome-wide level of significance ($p < 5 \times 10^{-8}$) in a large meta-analysis of the Global Lipids Genetics Consortium [24]. We extract beta-coefficients and standard errors of genetic associations for the 150 genetic variants and the 118 available metabolites. Next, we compute the genetic correlation structure between metabolites based on the $n = 150$ instrumental variables and exclude at random one of each pair of metabolites that are in stronger correlation than $|r| > 0.99$. For the simulation study each risk factor is scaled to have unit variance so all risk factors have an equal prior chance of being selected. Our final data-set β_X for the simulation study comprises associations of $d = 92$ metabolites measured on $n = 150$ genetic variants. This allows us to investigate risk factor selection for a realistic genetic correlation structure between metabolites (Supplementary Figure 2) and distribution of the regression coefficients.

Data pre-processing for blood cell traits for simulation

As a secondary data resource, we use publicly available summary data from the GWAS catalog `https://www.ebi.ac.uk/gwas/` on 36 blood cell traits measured on nearly 175 000 participants [21]. Using all genetic variants that

were genome-wide significant for any blood cell trait we have $n = 2667$ genetic variants as instrumental variables. There were eight pairs of blood cell traits with genetic correlation > 0.99. After removing three composite traits (sum of eutrophil and eosinophil counts, granulocyte count, and sum of basophil and neutrophil counts) from further analysis, there was no pair of blood cell traits with greater genetic correlation than 0.99. The respective correlation matrix is shown in Supplementary Figure 11. The final dataset used for the simulation consists of $d = 33$ blood cell traits as potential risk factors measured on $n = 2667$ genetic variants (pruned at $r^2 < 0.8$). For the simulation study each risk factor is scaled to have unit variance so all risk factors have an equal prior chance of being selected. We consider all $d = 33$ risk factors jointly for the simulation and consequently the simulation study has a realistic correlation structure between genetic associations of various blood cell traits (Supplementary Figure 11) and a realistic distribution of regression coefficients.

Data pre-processing and analysis for applied example of age-related macular degeneration

In the applied example we demonstrate how MR-BMA can be used to select metabolites as causal risk factors for age-related macular degeneration (AMD). As risk factors we consider a range of circulating metabolites measured by NMR spectroscopy [20]. We use the same lipid-related genetic variants as in the simulation study. We restrict the risk factor space to include only lipoprotein measurements on total cholesterol content, triglyc-

eride content, and particle diameter. For the various fatty acid measurements, we only included total fatty acids. Other lipid characteristics were highly correlated with the selected lipid measurements and including all of the lipid measurements would introduce multi-collinearity (RF2). As a next step we excluded all metabolite measures that did not have a single genetic variant that is genome-wide significant to meet the relevance criterion RF1. None of the remaining $d = 30$ metabolite measures have correlations in their genetic associations of $|r| > 0.985$ (Supplementary Figure 2). Genetic associations with the outcome are taken from the latest GWAS meta-analysis on AMD [25] including $16,144$ patients and $17,832$ controls which is available from `http://csg.sph.umich.edu/abecasis/public/amd2015/`. To synchronise the genetic data on the metabolite risk factors and the AMD outcome, we match the effect alleles and we remove two genetic variants missing in the AMD data, so that the overall analysis includes $n = 148$ variants. Finally, we use the Ensembl Variant Effect Predictor [26] to annotate the genetic variants to the gene that is most likely affected.

We run MR-BMA including all $n = 148$ available genetic variants on the $d = 30$ metabolite associations using $p = 0.1$ as prior probability, $\sigma^2 = 0.25$ as prior variance, a maximum model size of 12 risk factors, and with 100,000 iterations in the shotgun stochastic search. To check the impact of the prior choice we first vary the prior probability (Supplementary Table 7) of selecting a risk factor from $p = 0.01$ to $p = 0.3$ reflecting 0.49 to 14.7 expected causal risk factors. This choice alters the posterior probabilities of various individual models, but the overall marginal inclusion probabilities of the risk factors are relatively stable. Finally, we vary the prior variance σ^2 from 0.01 to 0.49,

which does not change the ranking (Supplementary Table 8).

References

[1] Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology* **178**, 1177–1184 (2013). URL `http://dx.doi.org/10.1093/aje/kwt084`.

[2] Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology* **37**, 658–665 (2013). URL `https://www.ncbi.nlm.nih.gov/pubmed/24114802`.

[3] Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine* **36**, 1783–1802 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28114746`.

[4] Burgess, S., Foley, C. N. & Zuber, V. Inferring Causal Relationships Between Risk Factors and Outcomes from Genome-Wide Association Study Data. *Annual Review of Genomics and Human Genetics* (2018). URL `https://doi.org/10.1146/annurev-genom-083117-021731`.

[5] Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine* **35**, 1880–1906 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/26661904`.

[6] Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology* **181**, 251–260 (2015). URL `https://www.ncbi.nlm.nih.gov/pubmed/25632051`.

[7] Thompson, S. G. & Sharp, S. J. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* **18**, 2693–708 (1999). URL `https://www.ncbi.nlm.nih.gov/pubmed/10521860`.

[8] Burgess, S. *et al.* Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. *Genetics* **207**, 481–487 (2017). URL http://www.genetics.org/content/207/2/481.

[9] Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian Randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology* dyy262–dyy262 (2018). URL http://dx.doi.org/10.1093/ije/dyy262.

[10] Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med* (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28960498.

[11] Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015). URL http://dx.doi.org/10.1093/ije/dyv080.

[12] Bowden, J., Smith, G. D., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40**, 304–314 (2016). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21965.

[13] Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17676998.

[14] Hans, C., Dobra, A. & West, M. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102**, 507–516 (2007). URL https://www.tandfonline.com/doi/abs/10.1198/016214507000000121.

[15] Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557–560 (2003). URL https://www.bmj.com/content/327/7414/557.

[16] Bowden, J. *et al.* Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology* (2018). URL https://doi.org/10.1093/ije/dyy258.

[17] Cook, R. D. Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174 (1979). URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481634.

[18] Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19**, 15–18 (1977). URL http://www.jstor.org/stable/1268249.

[19] Maddala, G. *Introduction to Econometrics* (Prentice Hall Professional Technical Reference, 1992), 2nd edn.

[20] Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications* **7** (2016). URL http://dx.doi.org/10.1038/ncomms11122.

[21] Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27863252.

[22] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Statist.* **32**, 407–499 (2004). URL https://projecteuclid.org:443/euclid.aos/1083178935.

[23] Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**, 1–22 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20808728.

[24] Consortium, G. L. G. Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274 EP – (2013). URL http://dx.doi.org/10.1038/ng.2797.

[25] Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics* **48**, 134 EP – (2015). URL http://dx.doi.org/10.1038/ng.3448.

[26] McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biology* **17**, 122 (2016). URL https://doi.org/10.1186/s13059-016-0974-4.