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

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

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

August 14, 2018

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 multivariable MR based on Bayesian model averaging (MR-BMA) that scales to high-throughput experiments and can select biomarker as causal risk factors for disease. 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.

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

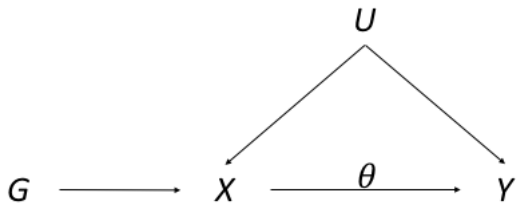

Figure 1: Directed acyclic graph of instrumental variable assumptions made in univariable Mendelian randomization. G = genetic variant(s), X = risk factor, Y = outcome, U = confounders, θ = causal effect of interest.

data. These biomarkers include proteins [3], blood cell traits [4], metabolites [5] or imaging phenotypes such as 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 summarized genetic data which is publicely available and allows to maximize the sample size.

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, triglyceride, phospholipid, and cholesterol esters, and additionally diameter of the lipoprotein particles. Apart from

lipoprotein measurements, this metabolite GWAS estimated genetic associations with amino acids, apolipoproteins, fatty and fluid acids, ketone bodies, and glycerides. This dataset also guides the design of our simulation study.

Results

Multivariable Mendelian randomizaton and risk factor selection

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, thus accounting for measured pleiotropy. The current implementation of multivariable MR is based on an inverse-variance weighted (IVW) linear regression in 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. Weights in these regression models 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 estimate from the multivariable MR represents 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]. 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.

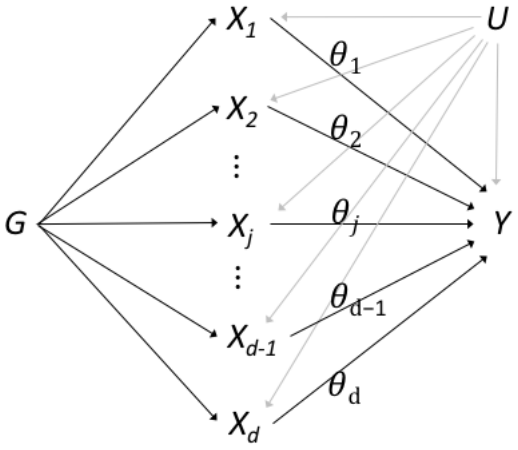

Figure 2: Directed acyclic graph of instrumental variable assumptions made in multivariable Mendelian randomization. G = genetic variant(s), X_j = risk factor j for $j = 1, \ldots, d$, Y = outcome, U = confounders, θ_j = causal effect of risk factor j.

To allow joint analysis of biomarkers from high-throughput experiments in multivariable MR we cast risk factor selection as variable selection in a weighted linear regression model. 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 aver-

aged causal effects ($MACE$) as the direct causal effect of a risk factor on an outcome. As we show in a simulation study on real biomarker data, MR-BMA enables sparse modeling and hence a better and more stable detection of the true causal risk factors.

Detection of invalid instruments

Invalid instruments may be detected as influential points or 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 [14]. 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 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) [15]. We illustrate the detection of influential points and outliers in the applied example.

Simulation results on NMR metabolite data

In a simulation study on a realistic data structure based on genetic associations with NMR metabolites [11] we compare the performance to detect true causal risk factors of the existing approach (Multivariable IVW regression), the Lasso [16], a penalised regression approach developed for high-dimensional regression models, our novel approach MR-BMA and the best model with the highest posterior probability from the Bayesian model selection. 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 [17]. We seek to evaluate two aspects of the methods: 1) how well can the competitors select the true causal risk factors (those with a non-zero causal effect), 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. Causal estimation is evaluated by calculating the mean squared error (MSE) of estimates, which captures both the bias and the variance properties of estimators. Each simulation scenario is repeated 1000 times.

Looking at a moderate set of $d = 12$ risk factors of which four risk factors are true causal ones, 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 using cross-validation and the Bayesian best model. Using Lasso with a weak penalisation (max) improves slightly over the standard IVW approach, which is showing the worst performance. The impact of the proportion of variance explained is shown in Supplementary Figure 1. Similar results are observed for setting B with eight true positive causal risk factors (Supplementary Figure 2). With respect to the MSE (Table 1), Lasso with the strongest penalty has the best performance for $R^2 = 0.1$, while for a larger proportion of variance explained MR-BMA has the lowest MSE. However, the MSE of Lasso with strong regularisation stays constant and does not decrease with increasing R^2 in contrast to all other approaches, suggesting that the low MSE is an artefact of the majority of the estimates being forced to zero. Again Lasso with weak regularisation gives similar results than the standard IVW, which performs worst of all methods with respect to the MSE. When looking at the distribution of causal estimates (Supplementary Figure 3 and 4) we find that IVW is the only method that gives unbiased estimates, although at the price of a high variance in the estimates. In contrast, estimates from Lasso with strong penalty are highly biased towards the null, but have very low variance. MR-BMA is a compromise between the strong Lasso penalty and the IVW estimate since it has a weaker bias towards the null than the Lasso but a much reduced variance compared to the IVW estimate.

When increasing the number of risk factors to $d = 92$ while keeping the number of true causal risk factors constant to four, the standard IVW methods 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 5 and 6. Despite being unbiased (Supplementary Figures 7 and 8), the variance of the IVW estimates is large and prohibits better performance. The Lasso provides sparse solutions with many of the causal estimates set to zero. This allows the Lasso a relative good performance at the beginning of the ROC curve, but its 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 again becomes apparent as the IVW method has a thousand times larger MSE than MR-BMA, which has the lowest MSE for all scenarios considered.

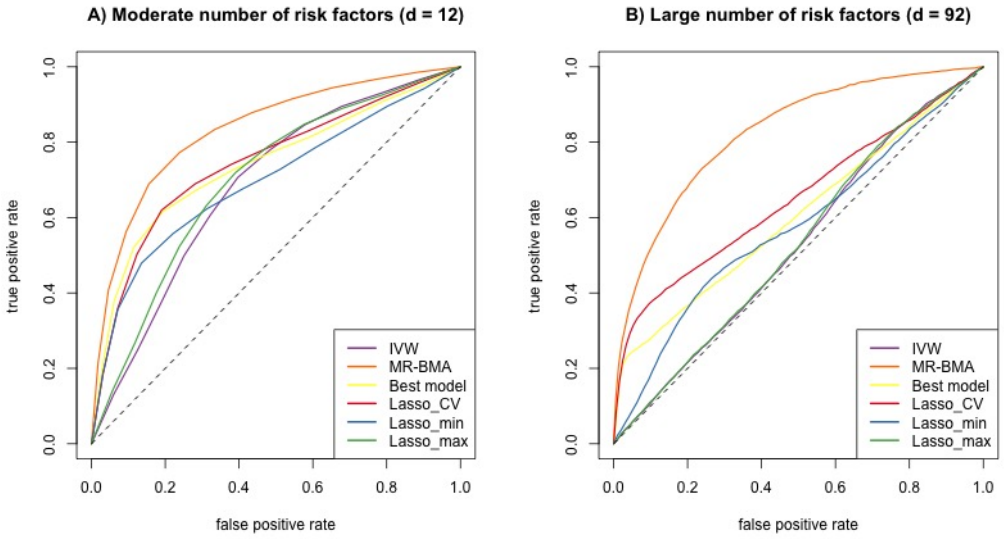

Figure 3: A) Receiver Operating Characteristic (ROC) curve for setting A including a moderate number of risk factors ($d = 12$) of which four have true positive effects. B) ROC curve for setting A including a large number of risk factors ($d = 92$) of which four have true positive effects. Proportion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

$d = 12$ risk factors	Setting A			Setting B		
R^2	0.1	0.3	0.5	0.1	0.3	0.5
IVW	0.742	0.180	0.077	0.592	0.158	0.071
Lasso (CV)	0.160	0.051	0.028	0.165	0.066	0.040
Lasso (min)	**0.028**	0.023	0.024	**0.051**	0.049	0.051
Lasso (max)	0.603	0.143	0.060	0.476	0.124	0.054
MR-BMA	0.039	**0.019**	**0.011**	0.059	**0.033**	**0.023**
Best model	0.072	0.032	0.017	0.100	0.050	0.030
$d = 92$ risk factors	Setting A			Setting B		
R^2	0.1	0.3	0.5	0.1	0.3	0.5
IVW	23.172	6.118	2.558	22.951	5.736	2.494
Lasso (CV)	0.053	0.014	0.008	0.051	0.015	0.011
Lasso (min)	0.345	0.090	0.038	0.341	0.084	0.036
Lasso (max)	18.912	4.991	2.083	18.729	4.673	2.028
MR-BMA	**0.005**	**0.004**	**0.003**	**0.009**	**0.008**	**0.007**
Best model	0.011	0.008	0.006	0.016	0.012	0.010

Table 1: Mean squared error (MSE) of the causal effect estimates from the competing methods. 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 [18]. 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. A recent multivariable MR analysis has shown that HDL-C may be a putative risk factor for AMD, while there was no evidence of a causal effect for LDL-C and triglycerides [19]. Here, we extend this analysis to consider not just three lipid measurements, but a wider range of $d = 49$ metabolite measurements as measured in the metabolite GWAS described earlier [11] for the same lipid-related instrumental variants as described previously. First, we prioritise and rank risk factors by their marginal inclusion probability (*MIP*) from MR-BMA. Secondly, we perform model diagnostics based on the best models with posterior probability > 0.01.

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, $MIP = 0.523$), all other risk factors have evidence less than $MIP < 0.25$. In order to check the model fit, we use the best individual models (Supplementary Table

1 B) with posterior probability > 0.01. 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, Supplementary Table 2) and Cook's distance (Figure 4 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 ($Cd = 0.295$ to $Cd = 0.565$) 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 [19, 20].

We repeat the analysis without the three influential and/or heterogeneous variants ($n = 145$), and report the ten risk factors with the largest marginal 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, $MIP = 0.677$), total cholesterol in large HDL particles (L.HDL.C, $MIP = 0.254$). 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 11 shows the scatterplots of β_X of each of these two risk factors individually against β_Y and their MACE estimates in red. We select the six individual models with a posterior probability > 0.01 to inspect the model fit (Supplementary Figures 12 and 13). This time, no observation has a consistently large q-statistic (Supplementary Table 5) or Cook's distance (Supplementary

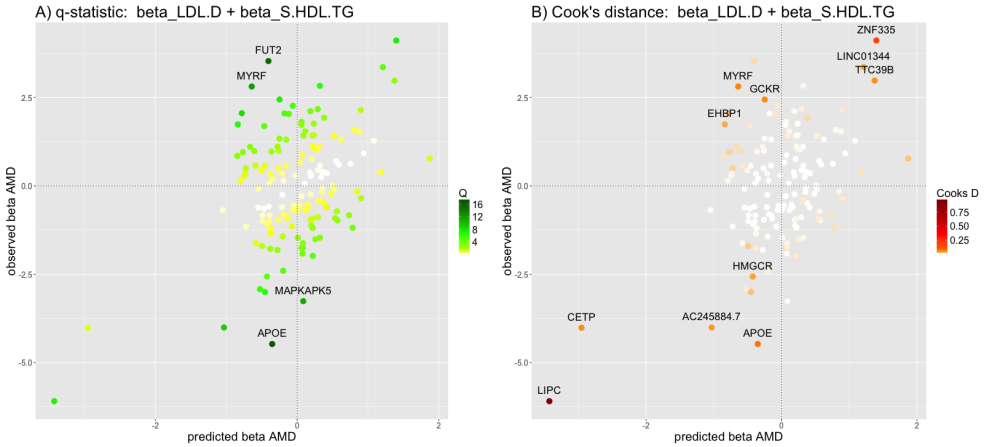

Figure 4: Diagnostic plot of the predicted associations with AMD based on the model including LDL.D, and S.HDL.TG (x-axis) against the observed associations with AMD (y-axis) including all $n = 148$ genetic variants. The color code shows: A) the q-statistic for outliers and B) Cook's distance for influential points. Any genetic variant with q-statistic > 10 or Cook's distance $> 4/n$ is marked by a label indicating the gene region. Note rs6859 in the *APOE* gene region with a q-statistic of 21.89 and rs261342 mapping to the *LIPC* gene region with a Cook's distance of 0.564.

Table 6) in any top models. 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.

These results confirm previous studies [19, 20] that identified HDL-C as a putative risk factor for AMD and draw the attention to extra-large and large HDL particles. 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

A) Model averaging

	Risk factor	Marginal inclusion probability (MIP)	Model-averaged causal estimate $\hat{\theta}_{\text{MACE}}$
1	XL.HDL.C	0.677	0.332
2	L.HDL.C	0.254	0.099
3	Glutamine	0.164	-0.055
4	Tyrosine	0.114	-0.034
5	HDL.D	0.085	0.023
6	XS.VLDL.TG	0.079	-0.018
7	Acetate	0.069	0.017
8	LDL.D	0.061	-0.014
9	IDL.TG	0.061	-0.011
10	S.VLDL.TG	0.059	-0.013

B) Individual models

	Risk factor(s)	Posterior probability (PP)	Model-specific causal estimates $\hat{\theta}_{\gamma}$
1	XL.HDL.C	0.068	0.509
2	L.HDL.C	0.034	0.384
3	Glutamine, L.HDL.C	0.014	-0.375,0.413
4	XL.HDL.C, XS.VLDL.TG	0.011	0.457,-0.181
5	IDL.TG, XL.HDL.C	0.011	-0.179,0.495
6	HDL.D	0.010	0.359
7	Tyrosine, XL.HDL.C	0.009	-0.275,0.532
8	Serum.C, XL.HDL.C	0.008	-0.183,0.573
9	Acetate, XL.HDL.C	0.008	0.272,0.494
10	Glutamine, XL.HDL.C	0.007	-0.263,0.514

Table 2: Ranking of risk factors for age-related macular degeneration (AMD): 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 model averaged causal effect 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, XS.VLDL.TG = Triglycerides in very small VLDL, XL.HDL.C = Total cholesterol in very large HDL

of variants underlining that the effect of large HDL particles is independent of the selection of instruments.

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. 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 [19,20] including three lipid measurements from the Global Lipids Genetics Consortium [17] have identified HDL-C as potential risk factor for AMD. Our new approach to multivariable MR refined this analysis using NMR metabolites as high-throughput risk factor set and confirmed HDL-C 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 [21]. 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 given 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.

When genetic variants are weak predictors for the risk factors, this can introduce weak instrument bias. In 2-sample MR, any bias due to weak instruments is towards the null and does not lead to inflated type 1 error rates [22]. 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 instruments. 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 0.98. 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 pleiotropy except for the measured risk factors. Pleiotropic variants can be detected as outliers to the model fit. Here we illustrate how to quantify outliers using the q-statistic. Outlier detection in the standard univariable MR approach can be performed by model averaging where different subsets of instruments are considered [23, 24], 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.

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).

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., Harbord, R., Sterne, J., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163 (2008).

[3] Sun, B. B. *et al.* Consequences of natural perturbations in the human plasma proteome. *bioRxiv* (2017). URL `https://www.biorxiv.org/content/early/2017/05/05/134551`. `https://www.biorxiv.org/content/early/2017/05/05/134551.full.pdf`.

[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 (2016).

[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. *Am J Epidemiol* **181**, 251–60 (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. http://www.genetics.org/content/207/2/481.full.pdf.

[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. *bioRxiv* (2018). URL https://www.biorxiv.org/content/early/2018/04/27/306209. https://www.biorxiv.org/content/early/2018/04/27/306209.full.pdf.

[14] 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. https://www.bmj.com/content/327/7414/557.full.pdf.

[15] 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. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481634.

[16] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Annals of Statistics* **32**, 407–451 (2004). URL <GotoISI>://WOS:000221411000001.

[17] 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.

[18] JM, S., J, C., WF, P., SH, A. & MC, N. The us twin study of age-related macular degeneration: Relative roles of genetic and environmental influences. *Archives of Ophthalmology* **123**, 321–327 (2005). URL http://dx.doi.org/10.1001/archopht.123.3.321.

[19] 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.

[20] 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. /oup/backfile/content_public/journal/ije/46/6/10.1093_ije_dyx189/2/dyx189.pdf.

[21] 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.

[22] 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. /oup/backfile/content_public/journal/aje/178/7/10.1093_aje_kwt084/1/kwt084.pdf.

[23] Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *International Journal of Epidemiology* (2017). Available online before print.

[24] 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 $\beta_{X_{ij}}$ and standard error $\text{se}(\beta_{X_{ij}})$ 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 β_{Y_i} and standard error $\text{se}(\beta_{Y_i})$ 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 β_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 randomisaton 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 β_{Y_i} are regressed on the genetic associations with the risk factor β_{X_i} [5]

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

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

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

Weights in these regression models are proportional to inverse of the variance of the genetic association with the outcome ($\text{se}(\beta_{Y_i})^{-2}$). This is to ensure that genetic variants having more precise association estimates receive more weight in the analysis. To account for heterogeneity in this equation, we can use a multiplicative random effects model, which increases the variance of the error terms by a multiplicative factor [7]. The same weighting can also be achieved by standardising the association estimates, by dividing β_{Y_i} and β_{X_i} by $\text{se}(\beta_{Y_i})$. In the following derivations, we assume that β_Y and β_X are

standardised, so that the variances of the ϵ_i terms are all 1. Our parameter of interest is the vector of regression coefficients $\theta = \{\theta_1, ..., \theta_d\}$. 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]. 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 alllows for measured pleiotropy [10] via any of the observed risk factors. 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].

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

$$\beta_Y \mid \beta_X, \theta, \tau \sim N(\beta_X \theta, \frac{1}{\tau}). \tag{3}$$

Following the *D2* prior specifications as introduced in [13], we use the following conjugate priors for the causal effects θ, the residual error ϵ, and the precision τ

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

where $\nu = diag(\sigma^2)$ 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 γ 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 γ encodes a specific regression model M_γ that includes the risk factors as indicated in γ. A model M_γ can include one or a combination of multiple risk factors. To evaluate the evidence of a specific model M_γ, we calculate the Bayes factor for model M_γ against the null model that does not include an intercept or any risk factor. The Bayes factor $BF(M_\gamma)$ has the

following closed form representation

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

where $\Theta = \Omega \beta_{X_\gamma}^t \beta_Y$ is the causal effect estimate and $\Omega = (\nu_\gamma^{-1} + \beta_{X_\gamma}^t \beta_{X_\gamma})^{-1}$ 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_γ of size k is defined as

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

The second important aspect is the prior for the variance of the risk

factors $\nu = diag(\sigma^2)$, where we assume that all risk factors have the same prior variance σ^2. Following [13] we set $\sigma^2 = 0.25$, but sensitivity of the results with respect to this prior should be investigated.

Posterior calculation and marginal inclusion probability of a risk factor

Let Γ be the space of all possible combinations of risk factors. The posterior probability (PP) of a model M_γ can be expressed by the prior probability (8) and the Bayes factor (6) of model M_γ as

$$PP(M_\gamma \mid \beta_Y, \beta_X) = \frac{p(M_\gamma)BF(M_\gamma)}{\sum_{\gamma\in\Gamma} p(M_\gamma)BF(M_\gamma)}. \tag{9}$$

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

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

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 [14] that evaluates all combinations of risk factors with a non-negligible contribution to the calibration factor $\sum_{\gamma\in\Gamma} p(M_\gamma)BF(M_\gamma)$ 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 $\hat{\theta}_\gamma$ of model M_γ as

$$\hat{\theta}_\gamma = \Omega\beta_{X_\gamma}^t\beta_Y = (\nu_\gamma^{-1} + \beta_{X_\gamma}^t\beta_{X_\gamma})^{-1}\beta_{X_\gamma}^t\beta_Y, \tag{11}$$

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

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

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 instruments

Invalid instruments may be detected as influential points or outliers with respect to the fit of a specific linear model M_γ. We recommend to check

the best individual models for outliers by visual inspection of the scatterplot of the predicted associations based on M_γ with the outcome $\hat{\beta}_Y = \beta_{X_\gamma}\hat{\theta}_\gamma$ against the actual observed observations 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

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

The individual element q_i measures the heterogeneity of a genetic variant i for a particular model M_γ. We refer to it as the q-statistic which we use 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 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 for a particular model M_γ we suggest to use Cook's distance [16]

$$Cd_i = \frac{q_i}{s^2 d} \frac{h_i}{(1-h_i)^2}, \tag{14}$$

where h_i is the ith diagonal element of the hat matrix $H = \beta_{X_\gamma}(\nu_\gamma^{-1} + \beta_{X_\gamma}^t \beta_{X_\gamma})^{-1}\beta_{X_\gamma}^t$, and $s^2 = \frac{1}{n-d}\epsilon^t\epsilon$ is the mean squared error of the regression model.

Simulation study on metabolite GWAS

To evaluate the performance of MR-BMA, we perform a simulation study using publicly-available summarized data on genetic associations with risk factors derived from a recent metabolite GWAS [17] as introduced earlier. 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 [18]. 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 $|r| > 0.99$. 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 and distribution

of the regression coefficients.

After taking genetic associations with the risk factors from the real dataset, we simulate genetic associations with the outcome β_Y based on a subset of risk factors, which we refer to as the 'true' risk factors. We investigate the following 12 different scenarios:

- Size of the data set: moderate ($d = 12$ metabolites selected at random) and large ($d = 92$ all metabolites available)
- Number of true risk factors: A) four risk factors have an effect of $\theta = 0.3$, the other risk factors have no effect. 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: $R^2 = 0.1, 0.3, 0.5$

We compare four different analysis methods:

- Multivariable inverse variance weighted (IVW) regression (equation 2) [19]
- Lasso as regularised regression [20]
- MR-BMA using marginal inclusion probabilities
- Bayesian best model selection using posterior probabilities of individual models

Lasso is a L1 regularised linear regression method which has been devised for variable selection in high-dimensional data. The regularisation parameter of Lasso is set to 0.1 (min, strong penalty) or 0.9 (max, weak penalty), where a penalty equal to 1 reflects the unpenalised IVW regression. Additionally, we use cross-validation (CV) to determine the penalty parameter. For the moderate risk factor space including 12 metabolites, the MR-BMA approach is performed using an exhaustive search of all possible models with a prior probability of a risk factor to be included set to $p = 0.5$, while for the large risk factor space 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 moderate risk factor space and around nine for the large risk factor space. The prior variance σ^2 is fixed to 0.25.

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 nuclear magnetic resonance (NMR) spectroscopy [17] from `http://computationalmedicine.fi/data#NMR_GWAS` and we use the same lipid-related instrumental variants as described previously. We restrict the risk factor space to include only lipoprotein measurements on total cholesterol content, triglyceride content, and particle diameter; for the various fatty acid measurements we only included total fatty acids. This results

in none of the $d = 49$ metabolite measures having correlations in their genetic associations of $|r| > 0.98$ (Supplementary Figure 9). Genetic associations with the outcome are taken from the latest large-scale GWAS meta-analysis on AMD [21] 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 [22] 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 $d = 49$ 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 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).

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`

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. /oup/backfile/content_public/journal/aje/178/7/10.1093_aje_kwt084/1/kwt084.pdf`.

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

[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).

[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).

[6] Burgess, S., Dudbridge, F. & Thompson, S. G. Re: "Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects". *American Journal of Epidemiology* **181**, 290–291 (2015).

[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. http://www.genetics.org/content/207/2/481.full.pdf.

[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. *bioRxiv* (2018). URL https://www.biorxiv.org/content/early/2018/04/27/306209. https://www.biorxiv.org/content/early/2018/04/27/306209.full.pdf.

[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. /oup/backfile/content_public/journal/ije/44/2/10.1093/ije/dyv080/2/dyv080.pdf.

[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. https://onlinelibrary.wiley.com/doi/pdf/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 <GotoISI>://WOS:000246859200015.

[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. https://www.bmj.com/content/327/7414/557.full.pdf.

[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. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481634.

[17] 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.

[18] 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.

[19] Burgess, S. & Thompson, S. G. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* **181**, 251–60 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25632051.

[20] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Annals of Statistics* **32**, 407–451 (2004). URL <GotoISI>://WOS:000221411000001.

[21] 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.

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