Mendelian Randomization Analysis of the Effect of Blood Lipid Fractions on the Risk of Type 2 Diabetes

Apostolos Gkatzionis[1]

[1]MRC Biostatistics Unit, University of Cambridge

Motivation

The aim of this analysis is to assess the causal effects of three blood lipid fractions (LDL-cholesterol, HDL-cholesterol and triglycerides) on the risk of suffering from type 2 diabetes. In parallel to that, we illustrate the use of the JAM-MR algorithm, a recently-proposed algorithm for pleiotropy-robust Mendelian randomization using summary data.

Data

We have only used a subset of the data provided for the MR Data Challenge. In particular, we have used the effect estimates, standard errors and univariate p-values for the association of 150 genetic variants with LDL-cholesterol, HDL-cholesterol, triglycerides and type 2 diabetes.

Analysis

To assess the effect of each of the blood lipid fractions on the risk of type 2 diabetes we performed a Mendelian randomization analysis using 150 genetic variants from the MR Data Challenge dataset. The genetic variants explained 6.12% of variation in HDL-cholesterol and 10.72% of variation in LDL-cholesterol measurements. For one of the variants in the dataset (rs261342 in the *LIPC* gene), the association with type 2 diabetes was missing so we excluded it from the analysis.

For each blood lipid trait, we performed three separate Mendelian randomization analyses. First, we used all 149 genetic variants. Second, we restricted ourselves only to the variants whose association with each trait passed the genome-wide significance threshold of 5×10^{-8}; this resulted in 64 genetic variants being used for HDL-cholesterol, 54 variants for LDL-cholesterol and 36 variants for triglycerides. Third, we refined our selection to include only those variants with genome-wide significant ($p < 5 \times 10^{-8}$)

associations with one trait but not with the other two traits. This resulted in 46 variants for HDL-cholesterol, 47 variants for LDL-cholesterol and 15 variants for triglycerides.

We then implemented a range of Mendelian randomization approaches to analyse the data. Th use of Mendelian randomization methods that account for the presence of pleiotropic variants is necessary because many variants are associated with multiple blood lipid traits: for example, 18 of the 64 SNPs associated with HDL-cholesterol at the genome-wide significance level were also associated with either LDL-cholesterol or triglycerides. Pleiotropy through different causal pathways is also possible.

We first performed univariate Mendelian randomization using an inverse-variance-weighted estimate for the causal effects of interest. Then, we implemented methods that account for the presence of pleiotropic variants in the dataset: we used MR-Egger, a weighted median and a weighted mode estimator. We also implemented the JAM-MR algorithm; more details about the algorithm can be found in the Technical Appendix and in Gkatzionis et al. (2019). Finally, we conducted a multivariate Mendelian randomization analysis, using all three traits simultaneously and all 149 genetic variants. We computed multivariate IVW estimates and estimates from multivariate MR-Egger regression.

Results

Table 1 summarizes the main results of our analysis. Results are reported as log-odds ratios of increase in type 2 diabetes risk per unit increase in the blood lipid concentration measurements. We list mean causal effect estimates and 95% confidence intervals obtained from each Mendelian randomization method for each of the three sets of genetic variants used. The causal effect estimates and associated 95% confidence intervals are also plotted in Figure 1

Our analysis suggests that LDL-cholesterol has a protective effect on the risk of type 2 diabetes. All the Mendelian randomization methods implemented for that trait indicated a negative causal effect, with the exception of mode-based estimation using all 149 genetic variants; standard errors for the mode-based method can be rather large when SNPs weakly associated with the risk factor are used, therefore its implementations using only the genome-wide significant variants are more reliable. For example, the JAM-MR causal effect estimate using only genome-wide significant genetic variants was -0.286 (95% confidence interval $(-0.415, -0.157)$), slightly larger than that obtained by the median and mode-based methods and fairly close to the estimate obtained from a multivariate Mendelian randomization analysis. Our results are consistent with the relevant literature, with studies showing that LDL-cholesterol reduction with statins results in a moderate increase in type 2 diabetes risk (Preiss et al., 2011; Swerdlow et al., 2015).

The majority of Mendelian randomization approaches yield a positive causal effect estimate for the association of triglyceride concentration with type 2 diabetes risk, with the only exception being MR-Egger. However, none of the reported associations is statistically significant, except for that obtained from the inverse variance weighted method using all 149 genetic variants. The results from the multivariate Mendelian randomization analysis

	HDL-cholesterol		LDL-cholesterol		Triglycerides	
	Mean	95%-CI	Mean	95%-CI	Mean	95%-CI
All 149 genetic variants						
IVW	-0.339	(−0.496, −0.182)	-0.183	(−0.339, −0.027)	0.248	(0.043, 0.453)
Egger	-0.345	(−0.572, −0.118)	-0.275	(−0.492, −0.058)	0.106	(−0.193, 0.405)
Median	-0.085	(−0.236, 0.066)	-0.204	(−0.342, −0.067)	0.168	(−0.054, 0.391)
Mode	0.045	(−2.341, 2.432)	-0.885	(−12.532, 10.761)	0.030	(−34.393, 34.454)
JAM-MR	-0.184	(−0.314, −0.054)	-0.279	(−0.399, −0.159)	0.114	(−0.093, 0.320)
GWAS-significant variants						
IVW	-0.382	(−0.575, −0.190)	-0.242	(−0.371, −0.113)	0.181	(−0.164, 0.525)
Egger	-0.029	(−0.374, 0.317)	-0.461	(−0.719, −0.203)	-0.646	(−1.354, 0.062)
Median	-0.070	(−0.228, 0.088)	-0.220	(−0.359, −0.081)	0.165	(−0.076, 0.406)
Mode	-0.006	(−0.165, 0.154)	-0.188	(−0.337, −0.039)	0.158	(−0.275, 0.591)
JAM-MR	-0.154	(−0.293, −0.015)	-0.286	(−0.415, −0.157)	0.216	(−0.018, 0.449)
Variants associated with only one trait						
IVW	-0.375	(−0.566, −0.183)	-0.166	(−0.277, −0.054)	0.037	(−0.288, 0.361)
Egger	-0.135	(−0.733, 0.463)	-0.283	(−0.513, −0.054)	-0.298	(−1.157, 0.561)
Median	-0.198	(−0.398, 0.002)	-0.200	(−0.343, −0.058)	0.178	(−0.157, 0.513)
Mode	-0.048	(−0.306, 0.210)	-0.194	(−0.346, −0.043)	0.199	(−0.301, 0.700)
JAM-MR	-0.203	(−0.391, −0.014)	-0.190	(−0.313, −0.067)	0.037	(−0.324, 0.398)
Multivariate Mendelian Randomization						
MV IVW	-0.280	(−0.466, −0.095)	-0.245	(−0.404, −0.085)	0.171	(−0.078, 0.420)
MV Egger	-0.309	(−0.541, −0.077)	-0.248	(−0.408, −0.087)	0.184	(−0.074, 0.441)

Table 1: Causal effect estimates and 95% confidence intervals for the effect of HDL-cholesteror, LDL-cholesterol and triglycerides on the risk of type 2 diabetes, estimated using a variety of Mendelian randomization approaches.

agree with the univariate methods and suggest no causal link between triglycerides and type 2 diabetes.

For HDL-cholesterol, the pattern of results is less clear. The standard inverse variance weighted estimates suggest a risk-decreasing effect on type 2 diabetes risk. Among the standard pleiotropy-adjusting methods, MR-Egger suggests a negative association when all genetic variants are used in the analysis, but does not identify a causal effect when only genome-wide significant variants are used. The weighted median and weighted mode estimators also do not suggest the existence of a causal effect. On the other hand, multivariate Mendelian randomization suggests a protective effect on HDL-cholesterol on type 2 diabetes. A possible explanation for this difference between multivariate and pleiotropy-robust Mendelian randomization methods would be the existence of variants which are associated with one of the blood lipid traits, not associated with the other two traits but have pleiotropic effects on type 2 diabetes through different biological mechanisms. The JAM-MR estimates lie in between those from the multivariate analysis and the pleiotropy-robust methods. The causal effect estimates are similar, regardless of which of the three sets of genetic variants we use (−0.184, −0.154 and −0.203). JAM-MR agrees with the

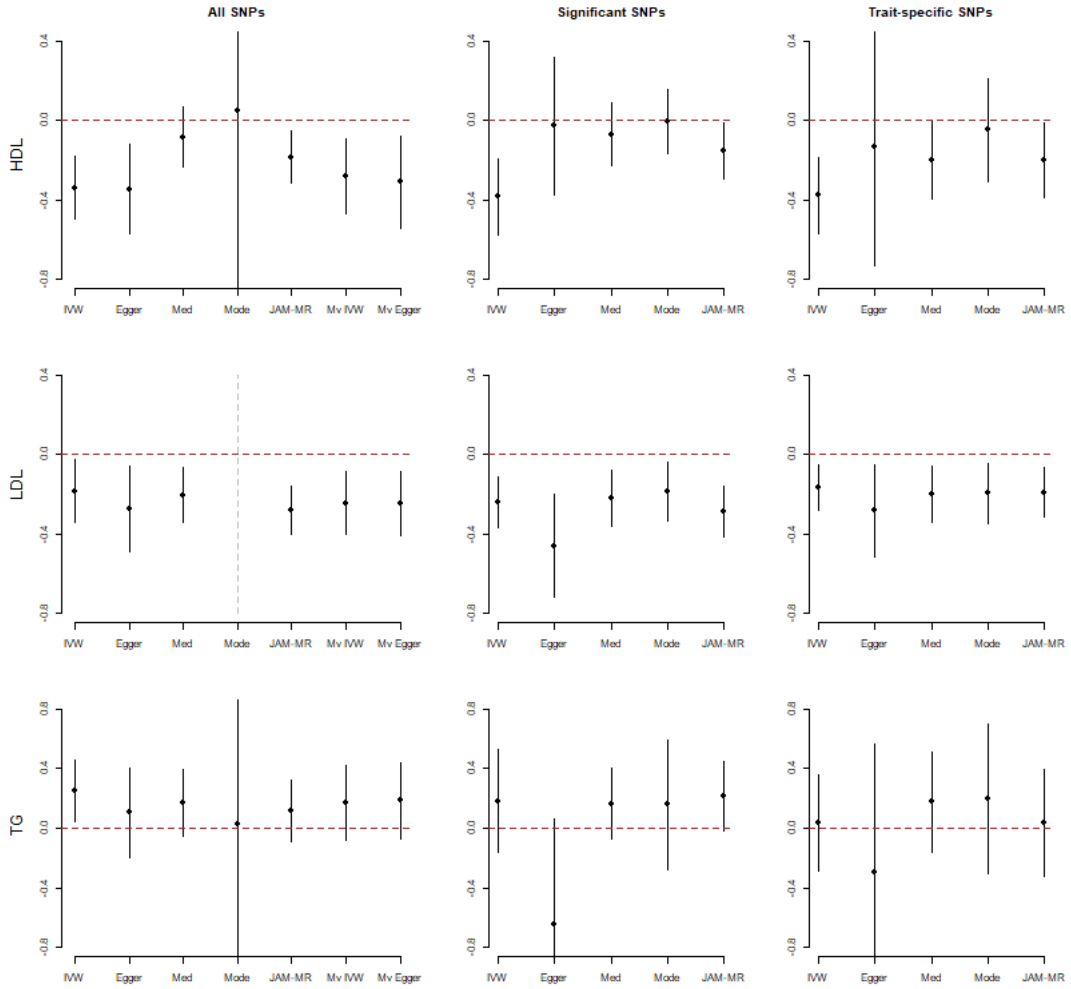

Figure 1: Causal effect estimates and 95% confidence intervals for the causal effects (log-odds ratios) of HDL-cholesterol, LDL-cholesterol and triglycerides on type 2 diabetes risk.

multivariate methods in suggesting a causal effect of HDL-cholesterol on type 2 diabetes in the negative direction.

The JAM-MR algorithm depends on a tuning parameter that determines the extent of heterogeneity penalization. Varying the value of that parameter can be a useful form of sensitivity analysis and can be used to explore the association between HDL-cholesterol and type 2 diabetes further. In Figure 2, we have plotted the causal effect estimates and corresponding standard errors obtained by the algorithm for a range of values of the tuning parameter. We have used the second set of genetic variants for this plot (SNPs associated with blood lipid traits at level $p < 5 \times 10^{-8}$), although similar plots were obtained using the other two sets of variants.

The plot for HDL-cholesterol gives an indication of how pleiotropy affects the analy-

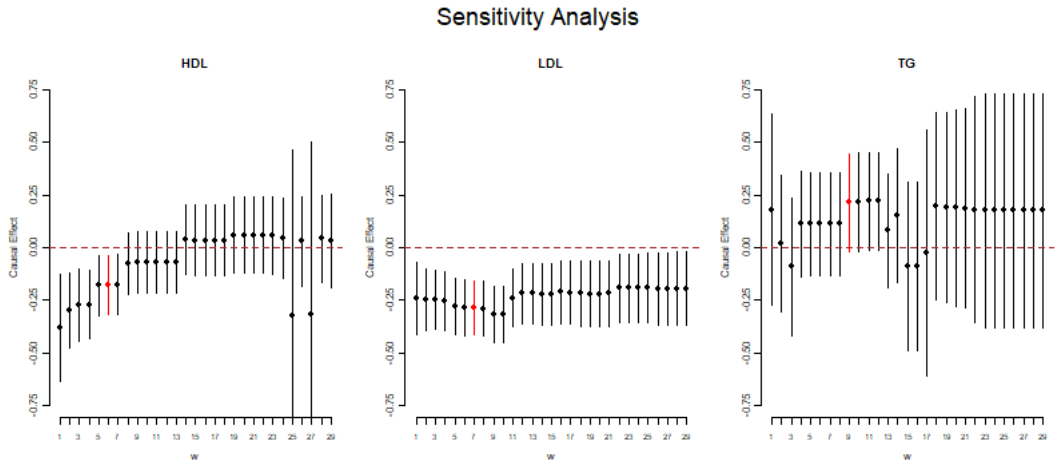

Figure 2: Causal effect estimates and 95% confidence intervals for the causal effects of blood lipids on type 2 diabetes risk, as estimated by JAM-MR with different degrees of pleiotropy penalization (different values of the tuning parameter). The runs with the smallest causal standard error are colored red. Significant SNPs only.

sis. The leftmost confidence interval practically corresponds to the IVW estimate, since no pleiotropy penalization is imposed. It suggests a protective effect of HDL-cholesterol on type 2 diabetes. As we impose stronger pleiotropy penalization this effect attenuates towards zero. Our selection criterion for tuning the algorithm aims to identify the largest collection of genetic variants with homogeneous effects (see the Technical Appendix). In this instance, the chosen JAM-MR implementation assigned posterior inclusion probability higher than 0.5 to 36 of the 64 genetic variants and still supported a weak preventive effect of high HDL-cholesterol on type 2 diabetes risk. However, the existence of a causal effect is not confirmed by all the genetic instruments. As we increased the value of the JAM-MR tuning parameter, the algorithm started consistently selecting a set of 13-17 genetic variants with very homogeneous effects that suggest no causal association.

To summarize, JAM-MR identified a set of 36 genetic variants with fairly homogeneous univariate causal effect estimates suggesting a weak protective effect of HDL-cholesterol on type 2 diabetes. Clustered within that set was a more homogeneous set of 17 genetic variants suggesting a null causal effect. A plausible explanation would be that these 17 genetic variants influence HDL-cholesterol concentration through the same biological mechanism, while the remaining 19 variants operate on a different biological process, although more research is needed to support this claim. If this is the case, we would expect JAM-MR to be able to separate these two sets of genetic variants more clearly if implemented using larger sample sizes.

The association between HDL-cholesterol and type 2 diabetes has been studied extensively in the literature, albeit with somewhat inconclusive results. Some Mendelian randomization analyses have indicated the existence of a protective effect (Fall et al., 2015; White et al., 2016) while others have suggested no causal relationship (Haase et al.,

2015). The potential for reverse causation has also been considered (Wang et al., 2018). Given that HDL-cholesterol is a composite risk factor, it is reasonable to expect that different genetic variants associated with it may act through different biological mechanisms. Our analysis shows that at least a subset of those variants that influence HDL-cholesterol levels are also associated with a reduction in type 2 diabetes risk. In order to disentangle the causal mechanism between those two traits, Mendelian randomization analyses considering more narrowly defined cardiovascular outcomes, such as those included in the MR Data Challenge dataset, may be helpful. The use of JAM-MR will hopefully help in understanding the complex relationship between HDL-cholesterol and type 2 diabetes.

Regarding LDL-cholesterol and triglycerides, Figure 2 is more consistent and suggests a protective effect of the former and no effect of the latter on type 2 diabetes risk.

The output from the JAM-MR algorithm is probabilistic: for each genetic variant, the algorithm returns a posterior probability of inclusion into the Mendelian randomization analysis. These inclusion probabilities can be plotted in a Manhattan plot, as illustrated in Figure 3 for the analyses using the genome-wide significant variants. Such Manhattan plots offer an additional way of visualizing the JAM-MR results. The runs depicted in Figure 3 assigned a posterior probability higher than 0.5 to 36 variants for HDL-cholesterol, 36 for LDL-cholesterol and 15 for triglycerides respectively. The identities of SNPs assigned high posterior probabilities are goven as part of the `R` code.

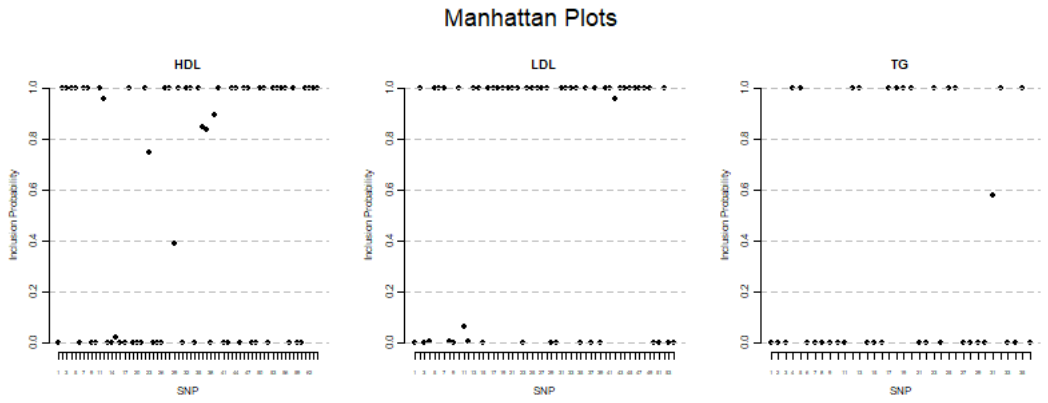

Figure 3: Manhattan plots of SNP inclusion probabilities for the Mendelian randomization analyses of the effects of blood lipids on type 2 diabetes risk. Significant SNPs only.

Technical Appendix

Contents

Here, we provide a brief overview of the JAM-MR algorithm. JAM-MR is an extension of the JAM algorithm (Newcombe et al., 2016), which was originally developed for fine-mapping densely genotyped regions. We first briefly review the JAM algorithm and then discuss how it can be adapted for Mendelian randomization. For more details, the reader is referred to Newcombe et al. (2016) and Gkatzionis et al. (2019).

The JAM Algorithm

The JAM algorithm performs Bayesian variable selection to identify genetic variants strongly associated with a trait of interest. Let X denote the trait and $G_1, \ldots, G_P$ denote a set of genetic variants potentially associated with X. JAM uses multivariate linear regression to model the trait. Given a subset γ of genetic variants and a set of individual measurements x_i, g_{ij}, JAM makes the modelling assumption

$$x_i | g_{ij}, \beta, \sigma^2, \gamma \sim N\left(\sum_{j \in \gamma} g_{ij}\beta_j \ \ , \ \ \sigma^2\right)$$

Let $x = (x_1, \ldots, x_N)$ be the vector of trait measurements and and $G = (g_{ij})$ be the genetic matrix. Let β_γ, G_γ be the sub-vector of β and the sub-matrix of G only for the SNPs in model γ. The likelihood associated with the JAM model is

$$p(x|\beta_\gamma, \sigma^2, \gamma, G) \propto \exp\left\{-\frac{1}{2\sigma^2}(x^T x - 2x^T G_\gamma \beta_\gamma + \beta_\gamma^T G_\gamma^T G_\gamma \beta_\gamma)\right\} \tag{1}$$

This expression depends on individual-level data only through the quantities $G_\gamma^T G_\gamma$, $G_\gamma^T x$ and $x^T x$. Of these quantities, $G_\gamma^T G_\gamma$ can be estimated from reference data and $G_\gamma^T x$ and $x^T x$ can be approximated by the GWAS summary statistics (see Newcombe et al., 2016, for more details). Therefore the JAM algorithm can be implemented using only GWAS summary statistics and does not require the availability of individual-level data.

The JAM likelihood (1) can be augmented with a set of priors. JAM uses conjugate (normal-inverse-gamma) g-priors for the model parameters, combined with a Beta-Binomial prior $p(\gamma)$ on the space of possible SNP configurations:

$$\begin{aligned} \beta_\gamma | \sigma^2, \gamma &\sim N\left(0, \sigma^2 \tau (G_\gamma^T G_\gamma)^{-1}\right) \\ \sigma^2 &\sim IG(a_X, b_X) \\ p(\gamma) &= \frac{B(a_\omega + P_\gamma, b_\omega + P - P_\gamma)}{B(a_\omega, b_\omega)} \end{aligned}$$

where $\tau = \max\{N, P^2\}$ and P_γ is the number of SNPs in model γ. Standard Bayesian inference yields the posterior $p(\beta_\gamma, \sigma^2, \gamma | x, G) \propto p(x|\beta_\gamma, \sigma^2, \gamma, G) p(\beta_\gamma|\sigma^2, \gamma) p(\sigma^2) p(\gamma)$. For

the purpose of variable selection, it is enough to restrict attention to the marginal model posterior

$$p(\gamma|x, G) = \int\int p(\beta_\gamma, \sigma^2, \gamma|x, G) d\beta_\gamma d\sigma^2 \tag{2}$$

The marginal model posterior can be used as part of a stochastic search algorithm. In particular, JAM implements a reversible-jump MCMC algorithm (Green, 1995) with addition, deletion and swapping of genetic variants as possible moves. This allows JAM to efficiently explore complex causal configurations among large numbers of genetic variants and select those robustly and independently associated with the trait.

The JAM-MR Algorithm

Suppose now that the trait X is to be used as a risk factor in a subsequent two-sample Mendelian randomization analysis in order to assess its effect on a disease outcome Y. We use θ to denote the $X - Y$ causal effect. The JAM-MR algorithm is an extension of JAM that performs SNP selection for this Mendelian randomization experiment.

Based on the instrumental variable assumptions, there are two requirements for the selection of genetic variants: first, the selected variants should be strongly associated with the risk factor X and second, they should not exhibit pleiotropic effects on the outcome Y. These two requirements can be combined by employing the framework of general Bayesian inference (Bissiri et al., 2016), an extension of traditional Bayesian inference where the likelihood is augmented with a loss function. For a dataset $\mathcal{D}$, a parameter vector λ and a prior distribution $\pi(\lambda)$, general Bayesian inference replaces the standard Bayesian updating scheme, $p(\lambda|\mathcal{D}) \propto \pi(\lambda)p(\mathcal{D}|\lambda)$, with a loss-function update of the form

$$p_\ell(\lambda|\mathcal{D}) \propto \pi(\lambda)p(\mathcal{D}|\lambda) \exp\{-w\ell(\mathcal{D}, \lambda)\} \tag{3}$$

where $\ell(\mathcal{D}|\lambda)$ denotes the loss function.

JAM-MR uses the JAM likelihood in place of $p(\mathcal{D}|\lambda)$ to prioritize the selection of variants strongly associated with the risk factor, and a heterogeneity-penalizing loss function to downweight variants which exhibit pleiotropic effects on the outcome. Let $\hat{\beta}_{X,j}$, $\hat{\beta}_{Y,j}$ be the univariate associations between genetic variant G_j and the risk factor and outcome respectively and $\hat{s}_{X,j}$, $\hat{s}_{Y,j}$ be the corresponding standard errors. Let $\hat{\theta}_j = \frac{\hat{\beta}_{Y,j}}{\hat{\beta}_{X,j}}$ be the univariate causal effect estimate for variant G_j and s.e. $\left(\hat{\theta}_j\right)$ its standard error. Finally, for model γ, let $\hat{\theta}_\gamma$ be the inverse variance weighted estimate of the causal effect θ obtained from all variants in that model. The loss function used by JAM-MR is

$$\ell(\gamma) = \frac{1}{P_\gamma - 1} \sum_{j \in \gamma} \left(\hat{\theta}_j - \hat{\theta}_\gamma\right)^2 \tag{4}$$

This is similar to the heterogeneity-penalizing loss function used in Burgess et al. (2018). A weighted version of (4) can be obtained by using an inverse-variance weight for each

term in the sum. Models containing zero or one genetic variants are excluded from the analysis.

If a model γ contains genetic variants with heterogeneous univariate causal effect estimates, the sum-of-squares term in the loss function will be large, the value of the loss-likelihood p_ℓ will be small and the model will be downweighted. On the other hand, if model γ contains genetic variants with homogeneous univariate causal effect estimates, the value of the loss function will be small and the model will be prioritized in JAM-MR's variable selection. Consequently, the algorithm tends to select the largest set of genetic variants with homogeneous causal effect estimates. We then implicitly assume that this set of variants corresponds to the valid instruments, similar to the plurality assumption made in Hartwig et al. (2016) and Burgess et al. (2018).

JAM-MR runs a stochastic search similar to JAM but using the loss-posterior (3)-(4) instead of the JAM likelihood (2). The stochastic search returns a set of most visited models and associated posterior probabilities. A causal effect estimate from each model is then computed using an inverse variance weighted formula, and an overall estimate is obtained by averaging across the different models. Standard errors for individual models are computed using a multiplicative random-effects formula, along with some additional adjustments (see Gkatzionis et al., 2019, for more details).

A final issue concerns the specification of the tuning parameter w in (3) that determines the strength of pleiotropy penalization. In the current version of the algorithm, the value of the tuning parameter is chosen according to a grid search after running the stochastic search for many different values. The selection criterion is that the resulting causal effect estimate should have the smallest standard error among all runs. Intuitively, this should happen when the model includes all the valid genetic variants and none of the pleiotropic ones. Given the model with all the valid SNPs, the inclusion of a pleiotropic SNP would increase the standard error due to the heterogeneity in univariate causal effect estimates. On the other hand, the removal of a valid SNP from the analysis would also increase the causal standard error, because the causal effect estimate would be based on fewer genetic variants. Nevertheless, exploring the performance of the algorithm for various values of the tuning parameter can be a useful form of sensitivity analysis, as illustrated in Figure 2.

Software

The data analysis was conducted in `R`. The code to implement the JAM-MR algorithm is available as part of the `R` package `R2BGLiMS`. For the implementation of the other Mendelian randomization methods, we used the `MendelianRandomization` package. The `R` code for our analysis is provided in a separate file.

References

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical*

Methodology), 78(5):1103–1130.

Burgess, S., Zuber, V., Gkatzionis, A., and Foley, C. N. (2018). 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*, 47(4):1242–1254.

Fall, T., Xie, W., Poon, W., Yaghootkar, H., Mägi, R., , Knowles, J. W., Lyssenko, V., Weedon, M., Frayling, T. M., and Ingelsson, E. (2015). Using genetic variants to assess the relationship between circulating lipids and type 2 diabetes. *Diabetes*, 64(7):2676–2684.

Gkatzionis, A., Burgess, S., Conti, D. V., and Newcombe, P. J. (2019). Bayesian variable selection with a pleiotropic loss function in Mendelian randomization. *bioRxiv*.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Haase, C. L., Tybjærg-Hansen, A., Nordestgaard, B. G., and Frikke-Schmidt, R. (2015). HDL cholesterol and risk of type 2 diabetes: A Mendelian randomization study. *Diabetes*, 64(9):3328–3333.

Hartwig, F. P., Davies, N. M., Hemani, G., and Davey Smith, G. (2016). Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International Journal of Epidemiology*, 45(6):1717–1726.

Newcombe, P. J., Conti, D. V., and Richardson, S. (2016). JAM: A scalable Bayesian framework for joint analysis of marginal SNP effects. *Genetic Epidemiology*, 40(3):188 – 201.

Preiss, D., Seshasai, S. R. K., Welsh, P., Murphy, S. A., Ho, J. E., Waters, D. D., DeMicco, D. A., Barter, P., Cannon, C. P., Sabatine, M. S., Braunwald, E., Kastelein, J. J. P., de Lemos, J. A., Blazing, M. A., Pedersen, T. R., Tikkanen, M. J., Sattar, N., and Ray, K. K. (2011). Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: A meta-analysis. *JAMA*, 305(24):2556–2564.

Swerdlow, D. I., Preiss, D., Kuchenbaecker, K. B., Holmes, M. V., Engmann, J. E. L., Shah, T., Sofat, R., and et al. (2015). HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet*, 385(9965):351 – 361.

Wang, N., Cheng, J., Ning, Z., Chen, Y., Han, B., Li, Q., Chen, C., Zhao, L., Xia, F., Lin, D., Guo, L., and Lu, Y. (2018). Type 2 diabetes and adiposity induce different lipid profile disorders: A Mendelian randomization analysis. *The Journal of Clinical Endocrinology and Metabolism*, 103(5):2016–2025.

White, J., Swerdlow, D. I., Preiss, D., Fairhurst-Hunter, Z., Keating, B. J., Asselbergs, F. W., Sattar, N., Humphries, S. E., Hingorani, A. D., and Holmes, M. V. (2016). Association of lipid fractions with risks for coronary artery disease and diabeteslipid fractions and risks of coronary artery disease and diabeteslipid fractions and risks of coronary artery disease and diabetes. *JAMA Cardiology*, 1(6):692–699.