Advanced Gene Mapping Course: Mendelian Randomization

Exercise Andrew DeWan, PhD, MPH

This exercise is designed to give you practical experience conducting a two-sample Mendelian randomization study. You can either use the online version of MR-base (https://www.mrbase.org/) or the accompanying R code to use TwoSampleMR.

You can search for the summary statistic data in MR-base directly, but another helpful resource for finding available statistic data is here: https://gwas.mrcieu.ac.uk

Part I:

You will be conducting an analysis to investigate the causal relationship between low density lipoprotein (LDL) and coronary heart disease (CHD) based on summary statistics from previously published GWAS data.

Exposure: Fasting LDL measurements from in 173,082 subjects and 2,437,752 genetic variants. Subjects are of European, East and South Asian and African ancestry.

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013 Nov;45(11):1274-1283. doi: 10.1038/ng.2797. Epub 2013 Oct 6. PMID: 24097068; PMCID: PMC3838666. GWAS ID: ieu-a-300

Outcome: CHD (e.g. myocardial infarction (MI), acute coronary syndrome, chronic stable angina, or coronary stenosis >50%) in 184,305 subjects (60,801 cases and 123,504 controls) and 9,455,779 genetic variants. Subjects are of European, East and South Asian, Hispanic and African ancestry.

Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015 Oct;47(10):1121-1130. doi: 10.1038/ng.3396. Epub 2015 Sep 7. PMID: 26343387; PMCID: PMC4589895. GWAS ID: ieu-a-7

1) Conduct an MR analysis of LDL and CHD. Studies can be searched by PubmedID (make sure PubmedID is checked) or specifying the GWAS ID. However, please note the following:

   A. For the exposure for this publication, use the larger set of subjects for this first analysis (N=173,082)
   B. For the exposure, use a p-value threshold of 5e10-8, LD Rsq = 0.001 and clumping distance of 10000kb. Also make sure "Perform Clumping" is checked if you're using MR-base
   C. For the outcome for this publication, use the trait denoted "Coronary heart disease"
   D. When running the MR analysis in MR-base you will want to allow LD proxies to be

selected for the outcome using a minimum Rsq of 0.8 and also allow for palindromic SNPs with a MAF threshold of 0.3. Make sure you set "Allele harmonization" to "Attempt to align strands for palindromic SNPs." In TwoSampleMR LD proxy parameters are set in the extract_outcome_data function and the allele harmonization option is set in the harmonise_data function with "action = 2".

E. Select the following methods:

a. Inverse variance weighted (NOTE: this is a random effects model)
b. MR Egger
c. Weighted Median

Questions:

1. How many variants are included in your genetic instrument for the exposure and how many are included in the outcome analysis? Of these, how many are proxies?

2. Based on the descriptions above, is the study used to define the IV appropriate for the outcome population?

3. Is there evidence of an association between LDL and CHD?

4. Is there evidence of heterogeneity in the genetic effects?

5. Is there evidence of pleiotropy?

6. How would you interpret the results of the three analyses together (i.e. IVW, MR Egger and Weighted Median)?

2) Re-run the analysis but for myocardial infarction (MI) using outcome data from the same publication (ieu-a-798).

Questions:

1. Is there evidence of an association between LDL and MI?

2. Can the association between LDL and CHD be explained by MI?

3) Feel free to explore associations with additional exposures such as HDL, BMI (you can use the Yengo et al. SNPs) or other exposures/outcomes of interest to you.

Part II:

Let's now see if we can validate the finding of an association between LDL and CHD by using different exposure data source and potentially dissect this signal to see if we pinpoint the features

of LDL that might be driving this signal. We will use metabolomics data that was generated in a sample of 24,925 individuals.

Kettunen, J., Demirkan, A., Würtz, P. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of *LPA*. *Nat Commun* **7,** 11122 (2016). https://doi.org/10.1038/ncomms11122.

Exposures: LDL.C, LDL.D, S.LDL.C, S.LDL.L, S.LDL.P, M.LDL.C, M.LDL.CE, M.LDL.L, M.LDL.P, M.LDL.PL, L.LDL.C, L.LDL.CE, L.LDL.FC, L.LDL.L, L.LDL.P, L.LDL.PL (16 metabolites)

Where S. = small, M. = medium and L. = large; .C = total cholesterol, .D = diameter, .L = total lipids, .P = concentration, .CE = cholesterol esters, .PL = phospholipids, .FC = free cholesterol

In MR-base, these can all be selected when you are on the "Choose Exposure" screen and selecting the "Metabolite level QTLs". You can then type in "LDL" in the analyte window and select each of these in the window that pops up. The most efficient way is to select all of the metabolites you're interested in and then run the MR analyses together. Before clicking off of this screen you will want to click on the "Select All" under Row Selection. This will allow you to run the analysis on all the SNPs for each metabolite. In TwoSampleMR these are brought in using data(metab_qtls) and then specifying the specific LDL variables above.

Use the same CHD outcome as you did for Part I (Nikpay PMID: 26343387, GWAS ID: ieu-a-7) using the full set of cases and controls (N=184,305).

1) Conduct an MR analysis as you did previously (NOTE: the TwoSampleMR code only specifies the IVW test for ease of summarizing the results but feel free to add the MR Egger and Weighted Median tests).

   Questions:

   1. For LDL.C, does the association between LDL and CHD validate the previous findings?

   2. Considering all the associations, do these results differentiate between the different characteristics of LDL (Please note: you will want to take into account the 16 association tests)?

3. What might be one explanation for the similarity between results for the different LDL characteristics?

4. Are there any concerns about heterogeneity or pleiotropy?

Part III:

Let's now look at the associations with VLDL metabolites using the same exposure and outcome data sources.

Exposures: 33 VLDL metabolites (Please note additional abbreviations: XS. = very small, XL. = very large, XXL. = extremely large; .TG = triglycerides)

1) Conduct an MR analysis as you did previously (NOTE: the TwoSampleMR code only specifies the IVW test for ease of summarizing the results but feel free to add the MR Egger and Weighted Median tests).

   Question:

   1. Considering all of the associations, are there any obvious trends in the results (Please note: you will want to take into account the 33 association tests)?

Answers

Questions:

1. How many variants are included in your genetic instrument for the exposure and how many are included in the outcome analysis? Of these, how many are proxies?

   There are 79 variants that surpass the p<5e-8 threshold for LDL in the exposure GWAS. Of these 77 are identified in the CHD outcome GWAS, 1 of which is a proxy.

2. Based on the descriptions above, is the study used to define the IV appropriate for the outcome population?

   They are generally well matched in terms of the population ancestries in the two studies, however, the outcome GWAS has subjects of Hispanic ancestry which could be a minor issue. This would be something to mention in the Discussion section of a manuscript. There is a subset of only European subjects for LDL but not for CHD, however, if you had access to the original data you could subset the subjects by ancestry to better match the exposure and outcome groups.

3. Is there evidence of an association between LDL and CHD?

   Yes, the IVW yields a beta = 0.4114 (p=1.626e-15) which corresponds to an OR of 1.51 (95% CI: 1.36 – 1.67) per SD increase in LDL.

4. Is there evidence of heterogeneity in the genetic effects?

   Yes, there is significant heterogeneity across effects of each SNP on CHD (p=2.822e-40) indicating that the random effects model is appropriate.

5. Is there evidence of pleiotropy?

   From the MR Egger regression there is no significant evidence of pleiotropy as the regression intercept is not significantly different from zero (p=0.118).

6. How would you interpret the results of the three analyses together (i.e. IVW, MR Egger and Weighted Median)?

   The IVW method (OR = 1.51, 95% CI: 1.36 – 1.67, p=1.626e-16), MR Egger (OR = 1.66, 95% CI: 1.42 – 1.93, p=1.086e-8) and Weighed Median (OR = 1.49, 95% CI: 1.36 – 1.63, p=1.962e-19) are relatively consistent meaning the causal effect estimate is likely to be between 1.49 and 1.66. There is no evidence that this estimate is influence by horizontal pleiotropy as the MR Egger intercept is not significant.

Questions:

1. Is there evidence of an association between LDL and MI?

Yes, IVW method provides significant evidence of an association between LDL and MI (OR = 1.48, 95% CI: 1.33 – 1.66, p=1.42e-12). The other MR measures of association are consistent with this estimate and there is again no evidence of horizontal pleiotropy.

2. Can the association between LDL and CHD be explained by MI?

We would need to test the other traits included in the CHD definition to see if they were associated with LDL or not and test for heterogeneity of the effects. However it is reassuring that the effect estimates are consistent between the larger CHD group and the smaller subgroup of subjects with MI.

3 Feel free to explore associations with additional exposures such as HDL, BM (you can use the Yengo et al. SNPs) or other exposures/outcomes of interest to you.

I'm more than happy to discuss additional results one-on-one or when we discuss the answers to this exercise.

Part II:

Questions:

1. For LDL.C, does the association between LDL and CHD validate the previous findings?

Yes, although the magnitude of the effect is slightly attenuated. The IVW yields a beta = 0.3665 (p=1.19e-07) which corresponds to an OR of 1.44. The previous OR estimate was 1.51 and the 95% CIs overlap.

2. Considering all the associations, do these results differentiate between the different characteristics of LDL (Please note: you will want to take into account the 16 association tests)?

All the associations yield statistically significant results, except for LDL diameter which doesn't meet the corrected significance threshold ($p<0.003125$). The remaining metabolites have statistically significant betas ranging from 0.3665 (LDL diameter) to 0.4709 (concentration of small LDL).

3. What might be one explanation for the similarity between results for the different LDL characteristics?

There is a high degree of overlap between the variants contained in the instrument variable

for each of the metabolites.

4. Are there any concerns about heterogeneity or pleiotropy?

There is significant heterogeneity across all the metabolites, except for the diameter of LDL. There is some evidence of pleiotropy among the large LDL metabolites, but not among any of the others.

Part III:

Question:

1. Considering all the associations, are there any obvious trends in the results (Please note: you will want to consider the 33 association tests)?

The results tend to be significant ($p<0.00015$) for the small and very small VLDL metabolites (except for the small VLDL triglycerides, $p=0.00019$), whereas the large, very large and extremely large LDL metabolites are non-significant (except for the cholesterol esters in large VLDL).