**Two-sample Mendelian randomization using summary genetic data – practical in MR Base**

Coronary heart disease (CHD) is one of the leading causes of death and morbidity in the world. Of the known risk factors, body mass index (BMI) is one of the most important. In prospective observational studies (Wormser et al Lancet 2011 377:1085), each standard deviation (SD) (~4.56 kg/m2) increase in BMI is associated with a relative risk of 1.23 (95% CI: 1.17–1.29) for coronary heart disease, after adjustment for conventional vascular risk factors, including smoking, cholesterol, diabetes, and blood pressure. However, whether the association reflects a causal effect of BMI on disease remains unclear.

We are going to use the R package that is released with MR Base, TwoSampleMR, to estimate whether there is a causal relationship between BMI and CHD. You can run a similar analysis in MR Base in your own time using the following link: <http://app.mrbase.org/>.

There is a good tutorial on how to use the TwoSampleMR R package that goes through some of the issues with these analyses in additional detail and I recommend you go through it before conducting these analyses using your own data: <https://mrcieu.github.io/TwoSampleMR/>.

There are two data files:

* BMIestimates-BMI\_SNPs.txt: this contains the SNP-exposure estimates for our analysis. The file contains the results for the 97 genome-wide significant SNPs identified in Locke et al 2015 (link to article: <https://www.nature.com/articles/nature14177>). They were selected from the file of all results from the Locke et al paper at the following link: <https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files> (“Download BMI EUR Ancestry GZIP”). Below is the first few rows of the file:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SNP | A1 | A2 | Freq1.Hapmap | b | se | p | N |
| rs1000940 | G | A | 0.225 | 0.0192 | 0.0034 | 1.28E-08 | 321836 |
| rs10132280 | A | C | 0.3333 | -0.023 | 0.0034 | 1.14E-11 | 321797 |
| rs1016287 | T | C | 0.325 | 0.0229 | 0.0034 | 2.25E-11 | 321969 |
| rs10182181 | A | G | 0.5 | -0.0307 | 0.0031 | 8.78E-24 | 321759 |
| rs10733682 | A | G | 0.425 | 0.0174 | 0.0031 | 1.83E-08 | 320727 |

* + SNP: this is the SNP identifier
  + A1: the effect allele
  + A2: the other (non-effect) allele
  + Freq1.Hapmap: the frequency of the A1 allele reported in HapMap
  + b: the beta (effect size) from the regression analysis of the SNP on BMI
  + se: the standard error from the regression analysis of the SNP on BMI
  + p: the P-value from the regression analysis of the SNP on BMI
  + N: the number of individuals included in the analysis (sample size)
* CHDestimates-BMI\_SNPs.txt: this contains the SNP-outcome estimates for our analysis. The file contains the results of the 97 genome-wide significant BMI associated SNPs for coronary heart disease identified in Nikpay et al 2015 (link to article: <https://www.nature.com/articles/ng.3396>). They were selected from the file of all results from the Nikpay et al paper at the following link: <http://www.cardiogramplusc4d.org/data-downloads/> (“CARDIoGRAMplusC4D 1000 Genomes-based GWAS - Additive [DOC 259,623KB]”). Below is the first few rows of the file:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| markername | chr | bp\_hg19 | effect\_allele | noneffect\_allele | effect\_allele\_freq | median\_info | beta | se\_dgc | p\_dgc | het\_pvalue |
| rs977747 | 1 | 47684677 | T | G | 0.449239 | 0.994 | 0.013896 | 0.009615 | 0.148375 | 0.95523 |
| rs657452 | 1 | 49589847 | A | G | 0.433876 | 0.999 | 0.00767 | 0.009384 | 0.413747 | 0.639827 |
| rs11583200 | 1 | 50559820 | C | T | 0.446987 | 1 | 0.00511 | 0.009533 | 0.591943 | 0.118195 |
| rs3101336 | 1 | 72751185 | C | T | 0.614647 | 1 | 0.011102 | 0.009766 | 0.255604 | 0.964553 |
| rs12566985 | 1 | 75002193 | G | A | 0.498752 | 0.99984 | -0.00239 | 0.009424 | 0.800033 | 0.508706 |

* + markername: this is the SNP identifier
  + chr: chromosome the SNP is on
  + bp\_hg19: chromosome position
  + effect\_allele: the effect allele
  + noneffect\_allele: the other (non-effect) allele
  + effect\_allele\_freq: frequency of the effect allele
  + median\_info: median imputation quality based on all participating studies
  + beta: the beta (effect size) from the regression analysis of the SNP on CHD
  + se\_dgc: the standard error (double genomic controlled) from the regression analysis of the SNP on CHD
  + p\_dgc: association P-value based on beta and se\_dgc
  + het\_pvalue: heterogeneity P-value testing the difference in effect size between studies

Let’s start by reading in our datasets and open the TwoSampleMR package:

# Install package (if these packages are already installed in your version of R then you can go down to # the ‘Load package’ section below)

install.packages("devtools")

library(devtools)

install\_github("MRCIEU/TwoSampleMR")

install.packages("plyr")

install.packages("ggplot2")

# Load package

library(TwoSampleMR)

library(plyr)

library(ggplot2)

# Set the working directory

setwd("C://Users/User/Desktop/PPU-GenEpi-main/Day4/")

# Load BMI (exposure) data

bmi\_exp\_data <- read\_exposure\_data(filename="BMIestimates-BMI\_SNPs.txt",

sep="\t", snp\_col="SNP", beta\_col="b", se\_col="se", pval\_col="p", eaf\_col="Freq1.Hapmap", effect\_allele\_col="A1", other\_allele\_col="A2", samplesize\_col="N")

bmi\_exp\_data$exposure <- "Body mass index"

# Look at the data and check the dimensions to ensure all SNPs are present

head(bmi\_exp\_data)

dim(bmi\_exp\_data)

# Load CHD (outcome) data

chd\_out\_data <- read\_outcome\_data(filename="CHDestimates-BMI\_SNPs.txt",

sep="\t", snp\_col="markername", beta\_col="beta", se\_col="se\_dgc", pval\_col="p\_dgc", eaf\_col="effect\_allele\_freq", effect\_allele\_col="effect\_allele", other\_allele\_col="noneffect\_allele")

chd\_out\_data$outcome <- "CHD"

# Look at the data and check the dimensions to ensure all SNPs are present

head(chd\_out\_data)

dim(chd\_out\_data)

# Harmonise the CHD and BMI datasets so that the effect alleles are the same.

# This syntax will flip the log odds ratio and effect alleles in the CARDIoGRAM dataset where the effect # alleles are different between CARDIoGRAMplusC4D and GIANT.

dat <- harmonise\_data(bmi\_exp\_data, chd\_out\_data)

# If you explore the dataset you'll notice that effect alleles and log odds ratios have been flipped in the # CHD dataset where the effect allele in the CHD dataset was different from the effect allele in the BMI # dataset

head(dat)

Now we are ready to start some analysis.

**A. Estimate the effect of genetically elevated BMI on CHD**

Estimate the effect of BMI on CHD using inverse-variance weighted (IVW) linear regression, simple median, weighted median and MR Egger regression. How do the results from each of the methods compare?

# Let's use the MR-Base R package to estimate the causal effect of BMI on CHD

res <- mr(dat,method\_list=c("mr\_ivw", "mr\_simple\_median",

"mr\_weighted\_median", "mr\_egger\_regression"))

# Results from the MR-Base package using various methods including MR-Egger and weighted median sensitivity analyses

res

#estimate odds ratio and 95% confidence interval

exp(res$b[1])

exp(res$b[1]-1.96\*res$se[1])

exp(res$b[1]+1.96\*res$se[1])

**B. Sensitivity analyses**

* + - 1. We discussed two tests for pleiotropy in the lectures; what were they? Is there any evidence for pleiotropy in our data?
      2. Can you think of any other way that we could test for pleiotropy?

# Test the intercept in MR-Egger to see if there is any evidence of directional pleiotropy

egg.int <- mr\_pleiotropy\_test(dat) # MR-Egger intercept test

egg.int

# Let's see if there is any heterogeneity between the Wald ratio estimates

het <- mr\_heterogeneity(dat, method\_list="mr\_ivw")

het

# Let's estimate the Wald ratio for each SNP to see if any particular SNP is having a big influence on the overall causal estimate

res\_single <- mr\_singlesnp(dat, all\_method=c("mr\_ivw",

"mr\_egger\_regression", "mr\_weighted\_median"))

res\_single

**C. Visualising the results**

1. Create a scatter plot of the SNP-CHD and SNP-BMI associations. Does the SNP-CHD association increase linearly as the SNP-BMI association increases? What could deviations from linearity mean? Are there any unusual data points?
2. Create a funnel plot of the results. Does the funnel plot look symmetric? What could asymmetry mean? Are there any outliers?
3. Create a forest plot of the results. Is there heterogeneity in the effects amongst SNPs? What could that indicate?

# Create a scatter plot of the SNP-CHD and SNP-BMI associations.

p1 <- mr\_scatter\_plot(res, dat)

p1

#save your plot using the png() function

png("scatter.png")

p1

dev.off()

# Create a funnel plot of the results.

res\_single <- mr\_singlesnp(dat)

p2 <- mr\_funnel\_plot(res\_single)

p2

#save your plot using the png() function

png("funnel.png")

p2

dev.off()

# 3. Create a forest plot of the results from the single SNP analysis in the previous section

p3 <- mr\_forest\_plot(res\_single)

p3

#save your plot using the png() function

png("forest.png")

p3

dev.off()

**D. Interpret the results**

1. Do you think BMI causes CHD? Consider these questions:

* 1. What is the odds ratio for coronary heart disease per unit increase in genetically elevated BMI?
  2. Are the genetic and observational effects directionally similar? Are they comparable?
  3. Are there any reasons we should be cautious with these results?