Introductory excercises in Mendelian randomisation

Gibran Hemani

26/08/2019

Objectives of this session

1. Gain familiarity with various R packages for performing two-sample MR
2. Reproduce several published MR analyses
3. Use range of sensitivity analyses to aid interpretation

We will use four different analyses for illustration:

- Urate on coronary heart disease
  - IVW vs other methods
- LDL cholesterol on alzheimer's disease
  - Outlier analysis
- Bi-directional education on intelligence
  - Steiger filtering
- Lipids on coronary heart disease
  - Multivariable MR

A note about software

In order to perform MR, you need to obtain the necessary data, put it into the format required for a particular package, and then analyse and interpret the results.

There are now several software packages available for MR analysis, for example

- TwoSampleMR package, developed in Bristol (part of the MR-Base project)
- Mendelian randomization R package, developed by Steve Burgess at Cambridge
- GSMR R package developed by Jian Yang and colleagues at University of Queensland
- RadialMR R package developed by Jack Bowden and Wes Spiller
- mrrobust stata package, developed by Tom Palmer, Wes Spiller, Neil Davies
- several more also arising now
- For two-stage least squares analysis with individual level data use the systemfit R package or ivreg in stata

For this practical we will be predominantly using the TwoSampleMR that can connect to the OpenGWAS database of GWAS summary data (https://gwas.mrcieu.ac.uk/) with functions for harmonising the data and analysing etc. It also can be used easily with several of the other packages.

You can find extended documentation on how to conduct various MR analyses here https://mrcieu.github.io/TwoSampleMR/

Sometimes the OpenGWAS servers have problems, so all the data has been pre-extracted here in case it is not available from the servers

```
load("pre_extracted.rdata")
```

Installation

If you are in a fresh environment that has not been setup for this practical, you will need to install the required packages. Do the following:

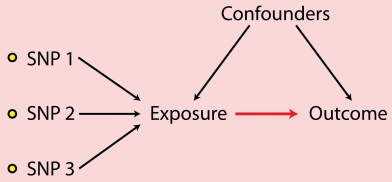
```
install.packages(c("devtools", "psych", "dplyr", "ggplot2", "plyr"))
devtools::install_github("MRCIEU/TwoSampleMR")
```

Basic workflow

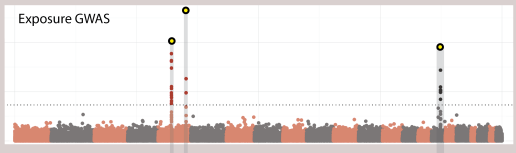
This schematic shows the steps required to perform two-sample MR

Performing Two Sample Mendelian Randomization

Objective: Infer the causal effect of the exposure on the outcome



1.



Description

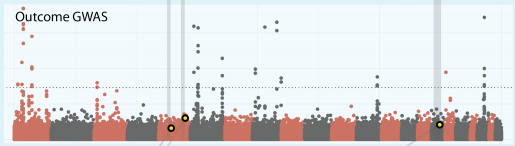
Define instruments: Obtain SNPs that are GWAS significant for the exposure. Ensure that they are independent.

Instruments can be defined from a variety of different sources.

R commands

```
extract_instruments()

read_exposure_data()

library(MRInstruments);
  data(gwas_catalog);
  data(aries_mqtl);
  data(gtex_eqtl);
  data(proteomic_qtls);
  data(metab_qtls);

clump_data()
```

2.



Get effects on outcome: Extract the instrument SNPs from the outcome GWAS. If they are not available, use LD proxies instead.

MR Base contains a large database of entire GWAS summary statistics.

```
extract_outcome_data()

read_outcome_data()
```

3.

	Exposure GWAS				Outcome GWAS			
SNP	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*
rs123456	0.132	A	G	0.28	0.022	A	G	0.26
rs234567	-0.485	G	T	0.41	**0.056**	**T**	**G**	**0.61**
rs345678	0.203	G	C	0.11	**-0.046**	**G**	**C**	**0.88**

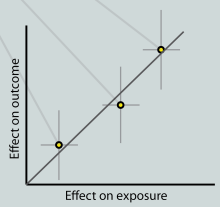
Harmonize effects

	Exposure GWAS				Outcome GWAS			
SNP	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*	*Effect*	*Effect allele*	*Other allele*	*Effect allele frequency*
rs123456	0.132	A	G	0.28	0.022	A	G	0.26
rs234567	-0.485	G	T	0.41	**-0.056**	**G**	**T**	**0.39**
rs345678	0.203	G	C	0.11	**0.046**	**G**	**C**	**0.12**

Harmonize effects: Ensure that the effect of the SNP on the exposure and the effect of the SNP on the outcome correspond to the same allele.

```
harmonise_data()
```

4.



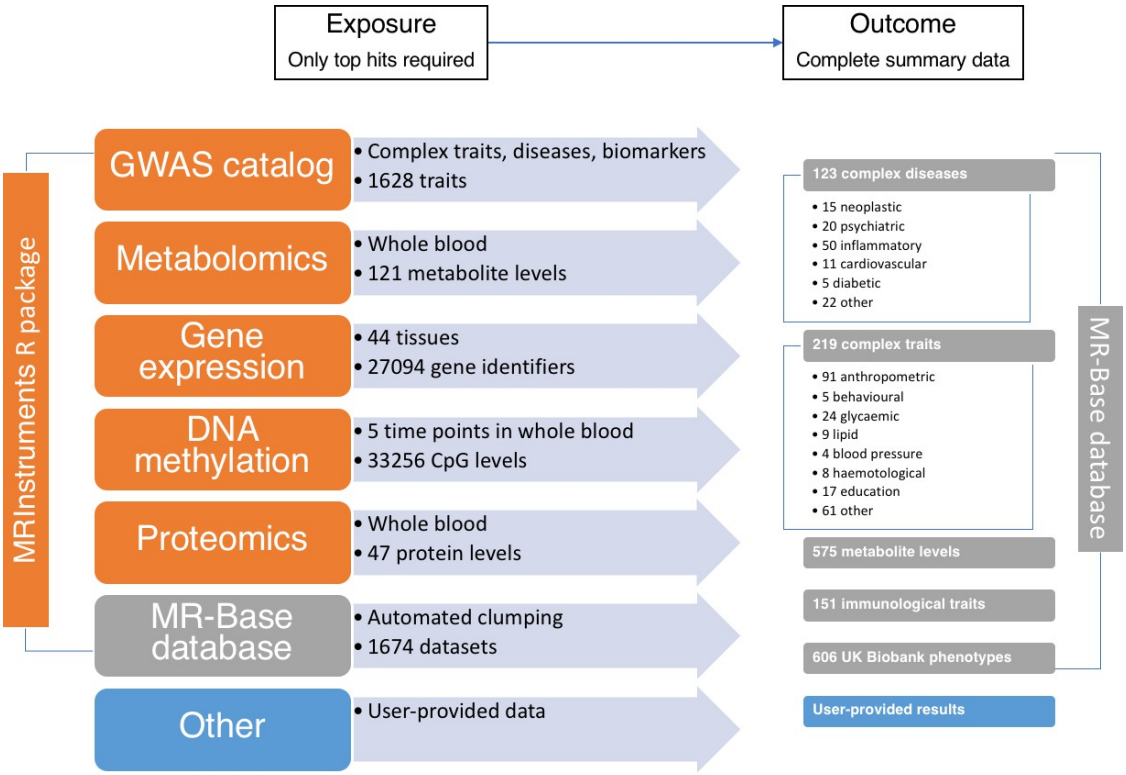
Perform analysis: Using the harmonized data, perform Mendelian randomization analyses and related sensitivity analyses.

The slope of the regression line corresponds to the causal effect of the exposure on the outcome

```
mr()

mr_singlesnp()

mr_leaveoneout()

mr_heterogeneity()

mr_steiger()

mr_pleiotropy_test()
```

Data available in OpenGWAS

We want to find the effect of an exposure on an outcome. An exposure can be analysed if instruments (i.e. GWAS hits) have been identified for it. Hence the only data required are the following for the top hits: rsid, effect size, standard error, effect allele. These are often recorded in various databases such as GWAS catalog etc. To test the effect of that exposure on the outcome, we need those same SNPs' effects on the outcome. There is no guarantee that they will have previously been GWAS hits, so a trait can only generally be analysed as an outcome if it has complete GWAS summary data available.



Exercises

1. The influence of urate levels in blood on coronary heart disease

This was the subject of an analysis by White et al 2016. First we need to obtain instruments for urate levels. A quick way to do this is to see if a suitably powered urate GWAS is available in OpenGWAS, and extract the LD-clumped top hits

```
library(TwoSampleMR)
ao <- available_outcomes()
subset(ao, grepl("Urate", trait))
```

The first result is the Kottgen et al 2013 GWAS using 110347 individuals. We can extract the top hits from this study

```
exposure_1 <- extract_instruments(outcomes = 1055)
head(exposure_1)
dim(exposure_1)
max(exposure_1$pval.exposure)
```

We have extracted 27 instruments for urate levels from this study.

Next we need to get the corresponding effects from a suitably powered coronary heart disease study

```
subset(ao, grepl("Coronary", trait))
```

The Nikpay et al 2015 study is very large (60801 cases), and is a genome-wide study with good imputation (9455779 SNPs)

```
outcome_1 <- extract_outcome_data(snps = exposure_1$SNP, outcome = 7)
head(outcome_1)
dim(outcome_1)
```

Next we have to harmonise the exposure and outcome data - meaning that the effect estimates are always on the same allele. e.g. we can see that the effect alleles are not always the same in the two studies:

```
merge(
    subset(exposure_1, select=c(SNP, effect_allele.exposure)),
    subset(outcome_1, select=c(SNP, effect_allele.outcome))
)
```

Harmonise:

```
dat_1 <- harmonise_data(exposure_1, outcome_1)
dim(dat_1)
table(dat_1$mr_keep)
```

DISCUSS: What has happened here - why are only 25 SNPs being retained for MR analysis?

We can now perform MR analysis on this harmonised dataset using the IVW method

```
res_1 <- mr(dat_1, method_list="mr_ivw")
res_1
```

Is there evidence for heterogeneity?

```
mr_heterogeneity(dat_1, method_list="mr_ivw")
```

It looks like there is substantial heterogeneity. Let's plot the results

```
mr_scatter_plot(res_1, dat_1)
mr_forest_plot(mr_singlesnp(dat_1, all_method="mr_ivw"))
```

We can try running a few sensitivity analyses

```
sens_1 <- mr(dat_1, method_list=c("mr_ivw", "mr_weighted_median",
"mr_egger_regression", "mr_weighted_mode"))
sens_1
mr_scatter_plot(sens_1, dat_1)
```

DISCUSS: How do we interpret the results now?

---

2. LDL cholesterol on Alzheimer's disease

We use this example to illustrate how outliers can make big influences on IVW analysis.

```
# The study ID for LDL cholesterol in the GLGC GWAS is 300
exposure_2 <- extract_instruments(300)

# Extract those SNPs from the IGAP Alzheimer's disease study (2013)
outcome_2 <- extract_outcome_data(exposure_2$SNP, 297)

# Harmonise
dat_2 <- harmonise_data(exposure_2, outcome_2)

res_2 <- mr(dat_2)
mr_scatter_plot(res_2, dat_2)
```

We can use the RadialMR R package to detect outliers

```
library(RadialMR)
dat_2_radial <- format_radial(BXG = dat_2$beta.exposure, BYG =
dat_2$beta.outcome, seBXG = dat_2$se.exposure, seBYG = dat_2$se.outcome,
RSID=dat_2$SNP)

ivwradial_2 <- ivw_radial(dat_2_radial, weights=1)
ivwradial_2$outliers
```

Remove the outliers and re-analyse

```
res_2_o <- mr(subset(dat_2, !SNP %in% ivwradial_2$outliers$SNP))
res_2
res_2_o
mr_scatter_plot(res_2_o, subset(dat_2, !SNP %in% ivwradial_2$outliers$SNP))
```

DISCUSS: How do the results compare before and after outlier removal? DISCUSS: What could bias the results aside from pleiotropy etc?

---

3. Education and intelligence

Much debate over the extent to which education influences intelligence and vice versa. The following two exercises reproduce (using slightly older data) the analyses performed by Anderson et al 2018. We can perform a bi-directional MR analysis, where we estimate the effects of education on intelligence, and then separately the effect of intelligence on education.

The MR-Base IDs for educational attainment and intelligence are `1001` and `UKB-a:196`, respectively

First do MR of education on intelligence

```
exposure_3a <- extract_instruments(1001)
outcome_3a <- extract_outcome_data(exposure_3a$SNP, "UKB-a:196")
dat_3a <- harmonise_data(exposure_3a, outcome_3a)
mr(dat_3a)
```

Now do the reverse, intelligence on education:

```
exposure_3b <- extract_instruments("UKB-a:196")
outcome_3b <- extract_outcome_data(exposure_3b$SNP, 1001)
dat_3b <- harmonise_data(exposure_3b, outcome_3b)
mr(dat_3b)
```

There are clearly very large effects in both directions. However, suppose that education is influeced by intelligence, and all the education instruments are actually just intelligence instruments - isn't a strong education-intelligence association exactly as we would expect? i.e. because we already know that the 'education SNPs' will have big effects on intelligence.

We can test the extent to which these SNPs are likely to be influencing education first and intelligence second, or vice versa. We do this by comparing the variance explained by the SNPs in the exposure against the outcome. We expect valid instruments to explain more variance in the exposure than the outcome.

```
dat_3a$units.outcome <- "SD"
dat_3a <- steiger_filtering(dat_3a)
dat_3b$units.exposure <- "SD"
dat_3b <- steiger_filtering(dat_3b)


# How many education SNPs influence education first
table(dat_3a$steiger_dir)

# How many intelligence SNPs influence intelligence first
table(dat_3b$steiger_dir)
```

Let's re-estimate the education-intelligence association, excluding SNPs that appear to influence intelligence first

```
mr(dat_3a)
mr(subset(dat_3a, steiger_dir))
```

We see that an effect remains, but it is almost halved from the original raw analysis.

---

4. Multivariable analysis of LDL, HDL and triglycerides on CHD

A major motivator for MR is to identify traits that we can intervene on for beneficial outcomes. The genetic influences on lipids are shared amongst the various subtypes, so it is difficult to gauge the specificity of the result from an MR analysis.

We can improve on single MR analyses by perform multivariable MR analysis, estimating the joint influences of several lipid traits on risk of coronary heart disease Burgess and Thompson 2015.

We will analyse LDL cholesterol (300), HDL cholesterol (299) and triglycerides (302) on CHD (7)

First let's look at the univariate analyses

```
exposure_4a <- extract_instruments(299)
outcome_4a <- extract_outcome_data(exposure_4a$SNP, 7)
dat_4a <- harmonise_data(exposure_4a, outcome_4a)

exposure_4b <- extract_instruments(300)
outcome_4b <- extract_outcome_data(exposure_4b$SNP, 7)
dat_4b <- harmonise_data(exposure_4b, outcome_4b)

exposure_4c <- extract_instruments(302)
outcome_4c <- extract_outcome_data(exposure_4c$SNP, 7)
dat_4c <- harmonise_data(exposure_4c, outcome_4c)

mr(dat_4a)
mr(dat_4b)
mr(dat_4c)
```

Higher LDL has a large effect on higher risk, but higher HDL looks like it might protect, and higher triglycerides might have higher risk. Let's see what multivariable analysis suggests

```
exposure_4d <- mv_extract_exposures(c(299,300,302))
outcome_4d <- extract_outcome_data(exposure_4d$SNP, 7)
dat_4d <- mv_harmonise_data(exposure_4d, outcome_4d)
mv_multiple(dat_4d)
```