# Testing instrument strength in two-sample MVMR estimation of lipid traits on AMD and Type 2 diabetes.

*Eleanor Sanderson and Jack Bowden*

*6th June 2019*

## Participants

Eleanor Sanderson and Jack Bowden, MRC Integrative Epidemiology Unit, Univerity of Bristol.

## Motivation

One of the key assumptions of any Mendelian randomisation (MR) analysis is that the exposures are strongly predicted by the set of SNPs used as instruments. In multivariable MR (MVMR) this assumption requires that each exposure is strongly predicted by the SNPs included conditional on the predicted value of the other exposures in the model. Many of the traits considered in the NMR data here are associated with highly overlapping groups of SNPs and therefore it is particularly important with data of this type to consider whether these exposures can be reliably predicted by the set of SNPs if multiple traits from this data are going to be included as exposures in an MVMR analysis.

Here I consider whether multiple traits in this dataset can be predicted at the same time. If a group of exposures can all be strongly predicted by the set of SNPs then, assuming the other instrumental variable assumptions are satisfied, it will be possible to estimate the direct effect of each exposure on an outcome. However, if some or all of the exposures are weakly predicted then any MR analysis including those exposures will be subject to weak instrument bias.

## Data

We use data of the effect sizes of each SNP on the 118 metabolites combined with the standard error of those SNP exposure associations (extracted from the GWAS results avaliable at http://www.computationalmedicine.fi/data#NMR_GWAS).(1) We also use data on the SNP associations with age related macular degeneration (AMD) from Fritsche et al 2016 (2). Finally we make use of the SNP effect sizes and standard error of these effect sizes on LDL, HDL and Triglicerides from Global Lipids consortium (3) and the data on type 2 Diabetes from the DIAGRAM consortium (4).

## Methods

Two sample summary data MR can be conducted using summary data estimates of SNP-exposure and SNP-outcome associations obtained from two independent but homogeneous study populations. For univariable

summary-data MR we can link the j'th SNP outcome association to the j'th SNP exposure association via the model;

$$\Gamma_j = \beta\pi_j$$

Where $\pi_j$ and $\Gamma_j$ represent the true association for SNP $G_j$ in G with the exposure and the outcome respectively. If the effect of each SNP on the exposure and outcome are estimated as;

$$x = \pi_0 + \pi_j G_j + \epsilon_{x,j}$$

and;

$$y = \Gamma_0 + \Gamma_j G_j + \epsilon_{y,j}$$

the Wald estimator $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\pi}_j$ is a consistent estimator for $\beta$. If the set of SNPs used as instruments is uncorrelated then taking an inverse variance weighted (IVW) average of the ratio estimates will yield an overall estimate for $\beta$ $\left(\hat{\beta}_{IVW}\right)$.

If we wish to estimate the effect of multiple exposures on an outcome we can use MVMR in the summary data setting to estimate the direct effect of each of the exposures included in the model on the outcome. This estimation involves regressing the SNP-outcome association on the SNP-exposure association for each of the exposures. For example, in the simplest case of two exposures the regression estimated is;

$$\hat{\Gamma}_j = \beta_1\hat{\gamma}_{1,j} + \beta_2\hat{\gamma}_{2,j} + v_j$$

Weighted by $\hat{\sigma}_{y,j}$, the variance of the estimated effect of SNP $j$ on the outcome. $\hat{\Gamma}_j$ is the estimated effect of SNP $j$ on the outcome and $\hat{\gamma}_{1,j}$ and $\hat{\gamma}_{2,j}$ are the estimated effects of SNP $j$ on exposures 1 and 2 respectively.(5,6)

The first assumption of MR analysis, that the instruments are strongly associated with the exposures, can be tested in a univariable two-sample MR through the mean F-statistic for the effect of the SNPs on the exposure. In a MVMR analysis it is no longer sufficient that this mean F-statistic is large but it is also necessary that the SNPs can strongly predict each exposure conditional on the other exposures. In analysis with individual level data this can be tested with a conditional F-statistic that tests the strength of the association of the SNPs with each exposure given the predicted values of the other exposures in the model.(2) We use an equivalent test for summary level data (see details in the technical appendix) to identify whether the SNPs here can predict a group of the NMR traits and the lipd fraction traits from the Global Lipids consortium. For each group of traits we then estimate the effect of those traits on a particular outcome, AMD for the NMR traits and type 2 diabetes for the Global Lipid consortium traits.

**1. Analysis NMR lipid traits.**

For this analysis we restrict the analysis to a limited subset of the avaliable traits. Therefore we consider the 10 traits with the largest mean F-statistic for the association between each SNP and the exposure for the SNPs that are individually strongly associated with that trait, we define SNPs as being associated with a trait if the individual F-statistic for that SNP-trait association is greater than 10. This restriction is equivalent to selecting the 10 traits with the smallest mean p-value within the group of SNPs strongly associated with that trait. From this set of traits we estimate an MVMR analysis using IVW for all 10 traits on AMD. We then investigate the conditional F-statstics for this group of traits to determine which set can be jointly predicted by the group of SNPs and re-estimate our MVMR analysis using only the group of traits that can be jointly strongly predicted by the SNPs as our exposures.

**2. Analysis lipid fraction traits from Global Lipids consortium and type 2 diabetes.**

For this analysis we considered the effect of the three lipid fraction traits, HDL, LDL and Trigycerides on type 2 diabetes. We again restrict the analysis to SNPs with an individual F-statistic greater than 10 for at least one of these traits. We estimate the conditional F-statistic for each trait to determine whether they can be strongly predicted, as a group, by the set of SNPs and then estimate the direct effect of each trait on type 2 diabetes.

For each MVMR analysis we estimate the results with and without an iteratively updated weight to account for the error in the SNP-exposure association.(7) IVW estimation assumes that there is no uncertainty in the SNP-exposure associations. By changing the weighting used in the analysis to account for the uncertainty in the SNP-exposure associations we can relax this assumption and obtain more accurate estimates of the effect of the exposures on the outcome.

## Results

**1. Analysis NMR lipid traits.**

Table 1 gives the results from a MVMR regression of AMD on the 10 exposures with the highest individuals F statstics in the data and simple weights. Many of the point estimates in this table are so large that it is clear that this analysis is not reliable.

Table 2 gives the Conditional F-statistic, individial F statistic for the full set of SNPs included in the MVMR and the individual F-staitistic for the SNPs that are strongly associated with each trait for each trait included in the MVMR analysis. Each exposure should be considered strongly predicted in the MVMR analysis if the conditional F-statistic is larger than the rule of thumb of 10. In this case the conditional F-statistics are all very small showing that this groups of traits cannot be jointly strongly predicted by these SNPs.

Examining these traits more closely we see that they fall into three groups - Fatty Acids (“CH2.in.FA” “Bis.DB.ratio” and “Bis.FA.ratio”), small HDL ( “S.HDL.P” and “S.HDL.L”) and very large HDL (“XL.HDL.TG”, “XL.HDL.P”, “XL.HDL.PL” , “XL.HDL.FC” and “XL.HDL.L”). Tables 3 - 5 give the conditional F statistic calculated within each group and show that it is not possible to jointly predict traits from the same group. Table 6 gives the same results including only one exposure from each group and shows

that once traits from different categories are included it is possible to jointly strongly predict these traits and therefore they can all be included in an MVMR analysis.

In Table 7 we additionally included Triglycerides in very large HDL, these results show that although the traits are now more weakly predicted each trait still has a conditional F statistic larger than 10 and so this set of traits could be strongly predicted in a MVMR estimation.

Table 8 gives the results from a MVMR estimation of the effect of this set of traits on AMD. In Table 9 these results have been updated using an iteratively updated IVW estimation (with 4 iterations). The results from this analysis show a potential effect of very large HDL and triglycerides in very large HDL on AMD, conditional on small HDL and Fatty acid levels. In this example updating the weights has changed the point estimates obtained from the analysis as there is a high level of uncertainty around the SNP exposure associations for this set of SNPs and exposures.

Table 1. MVMR estimation of AMD on the 10 most strongly predicted NMR traits.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| CH2.in.FA | -0.2342 | 0.4346 | -0.5387 | 0.5967 |
| S.HDL.P | 2.7609 | 1.6340 | 1.6896 | 0.1084 |
| S.HDL.L | -3.4312 | 1.7168 | -1.9986 | 0.0610 |
| XL.HDL.TG | -0.8570 | 0.3019 | -2.8387 | 0.0109 |
| Bis.FA.ratio | 1.4840 | 1.1366 | 1.3056 | 0.2081 |
| XL.HDL.P | -4.7100 | 1.3096 | -3.5966 | 0.0021 |
| Bis.DB.ratio | -1.3887 | 1.1263 | -1.2329 | 0.2335 |
| XL.HDL.PL | -2.5502 | 1.5646 | -1.6299 | 0.1205 |
| XL.HDL.FC | 2.6429 | 2.2429 | 1.1784 | 0.2540 |
| XL.HDL.L | 5.2654 | 2.8497 | 1.8477 | 0.0812 |

Table 2. Individual and Conditional F-statistics for the 10 MNR tratits.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| CH2.in.FA | 0.413 | 5.319 | 29 | 91.646 | 1 |
| S.HDL.P | 0.075 | 11.993 | 29 | 91.482 | 3 |
| S.HDL.L | 0.069 | 11.817 | 29 | 69.235 | 4 |
| XL.HDL.TG | 0.107 | 29.579 | 29 | 57.791 | 14 |
| Bis.FA.ratio | 0.092 | 9.427 | 29 | 53.685 | 4 |
| XL.HDL.P | 0.096 | 17.175 | 29 | 52.634 | 8 |
| Bis.DB.ratio | 0.097 | 10.504 | 29 | 51.468 | 5 |
| XL.HDL.PL | 0.094 | 22.856 | 29 | 46.828 | 13 |
| XL.HDL.FC | 0.036 | 18.301 | 29 | 42.869 | 11 |
| XL.HDL.L | 0.034 | 19.517 | 29 | 42.300 | 12 |

Table 3. Individual and conditional F-statistics for the Fatty Acid traits.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| CH2.in.FA | 0.286 | 22.019 | 5 | 91.646 | 1 |
| Bis.DB.ratio | 0.063 | 51.468 | 5 | 51.468 | 5 |
| Bis.FA.ratio | 0.067 | 44.830 | 5 | 53.685 | 4 |

Table 4. Individual and conditional F-statstics for the small HDL traits.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| S.HDL.P | 0.038 | 70.977 | 4 | 91.482 | 3 |
| S.HDL.L | 0.038 | 69.235 | 4 | 69.235 | 4 |

Table 5. Individual and conditional F-statistics for the very large HDL traits.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| XL.HDL.TG | 0.120 | 33.567 | 25 | 57.791 | 14 |
| XL.HDL.P | 0.124 | 19.685 | 25 | 52.634 | 8 |
| XL.HDL.PL | 0.135 | 26.002 | 25 | 46.828 | 13 |
| XL.HDL.FC | 0.058 | 21.036 | 25 | 42.869 | 11 |
| XL.HDL.L | 0.042 | 22.344 | 25 | 42.300 | 12 |

Table 6. Individual and conditional F-statistics for strongest predicted trait from each group.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| Bis.DB.ratio | 24.491 | 20.088 | 14 | 51.468 | 5 |
| S.HDL.P | 25.123 | 21.533 | 14 | 91.482 | 3 |
| XL.HDL.P | 33.014 | 32.116 | 14 | 52.634 | 8 |

Table 7. Individual and conditional F-statistics for strongest predicted trait from each group and Triglycerides in very large HDL.

| | Conditional F Stat | MVMR Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| Bis.DB.ratio | 13.467 | 12.235 | 24 | 51.468 | 5 |
| S.HDL.P | 17.706 | 13.620 | 24 | 91.482 | 3 |
| XL.HDL.P | 9.656 | 19.455 | 24 | 52.634 | 8 |
| XL.HDL.TG | 11.934 | 34.886 | 24 | 57.791 | 14 |

Table 8. MVMR for the effect of 4 jointly strongly predicted traits on AMD, estimated using IVW.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Bis.DB.ratio | 0.1588 | 0.1694 | 0.9375 | 0.3603 |
| S.HDL.P | -0.0418 | 0.2461 | -0.1698 | 0.8670 |
| XL.HDL.P | 0.6725 | 0.3262 | 2.0613 | 0.0532 |
| XL.HDL.TG | -0.6515 | 0.2454 | -2.6552 | 0.0156 |

Table 9. MVMR for the effect of 4 jointly strongly predicted traits on AMD, estimated using iteratively updated IVW.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Bis.DB.ratio | 0.1941 | 0.1979 | 0.9809 | 0.3390 |
| S.HDL.P | 0.0310 | 0.2129 | 0.1457 | 0.8857 |
| XL.HDL.P | 0.6302 | 0.2635 | 2.3920 | 0.0273 |
| XL.HDL.TG | -0.5383 | 0.2320 | -2.3205 | 0.0316 |

**2. Analysis lipid fraction traits from Global Lipids consortium and type 2 diabetes.**

Table 10 gives the individual and conditional F-statistics for the three lipid fraction traits from the Global Lipids consortium. These results show that all three traits can be strongly preicted by this set of SNPs, however they also highlight that even in this case the exposures are less strongly predicted jointly than they are individually.

Table 11 gives the results for an MVMR analysis of these traits on risk of type 2 diabetes using the standard IVW estimate and Table 12 gives results from the same analysis with iteratively updated weights to account for the uncertianty in the SNP-exposure association. These results show that both HDL and LDL fractions appear to have a direct negative effect on type 2 diabetes risk once the other lipid fractions have been controlled for. In this case updating the IWV estimate does not have a large effect on the results obtained as the SNP exposure associations are estimated more precisely than in the previous example.

Table 10. Individual and Conditional F statistics for LDL, HDL and Triglycerides.

| | Conditional F Stat | Mean Individual F stat | No of SNPs | Individual F stat | No of SNPs |
|---|---|---|---|---|---|
| ldl | 29.898 | 66.526 | 146 | 127.368 | 75 |
| hdl | 23.434 | 105.895 | 146 | 166.759 | 92 |
| tg | 18.256 | 44.650 | 146 | 87.330 | 72 |

Table 11. MVMR estimation of the effect of Lipid fractions on Type 2 Diabetes, estimated using IVW.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| hdl | -0.2273 | 0.0766 | -2.9679 | 0.0035 |
| ldl | -0.3150 | 0.0687 | -4.5833 | 0.0000 |
| tg | 0.1144 | 0.1113 | 1.0286 | 0.3054 |

Table 12. MVMR estimation of the effect of Lipid fractions on Type 2 Diabetes, estimated using IVW with iteratively updated weights.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| hdl | -0.2229 | 0.0764 | -2.9179 | 0.0041 |
| ldl | -0.3256 | 0.0684 | -4.7568 | 0.0000 |
| tg | 0.1323 | 0.1111 | 1.1913 | 0.2355 |

## Technical Appendix

In two-sample MVMR testing whether the SNPs can explain variation in one exposure conditional on the other exposures has been shown elsewhere to be equivalent to testing (in an example with two exposures) whether the model;

$$X_1 = \delta_0 + \delta_1 X_2 + u_1$$
$$X_2 = \pi_0 + \pi_1 G + u_2$$

is overidentified.(8) The analogous estimation for two-sample MR is;

$$\hat{\pi}_{i,j} = \delta_1 \hat{\pi}_{2,j} + \epsilon \tag{1}$$

Overidentification in this model can be tested using a modified version of Cochran's Q statistic;

$$Q_{x_1} = \sum_{j=1}^{L} \left( \frac{1}{\sigma^2_{x_{1,j}}} \right) \left( \hat{\pi}_{i,j} - \tilde{\delta} \hat{\pi}_{2,j} \right)^2 \tag{2}$$

Where $\sigma^2_{x_{1,j}} = \sigma^2_{1,j} + \tilde{\delta}^2 \sigma^2_{2,j} - 2\tilde{\delta}\sigma_{12,j}$. $\sigma^2_{1,j}$ is the variance of $\hat{\pi}_{1,j}$, $\sigma^2_{2,j}$ is the variance of $\hat{\pi}_{2,j}$, $\sigma_{12,j}$ is the covariance of $\hat{\pi}_{1,j}$ and $\hat{\pi}_{2,j}$ and $\tilde{\delta}$ is an efficient estimator of $\delta$. We therefore propose using this test to identify weak instruments in two-sample MVMR. Under the null hypothesis that the instruments do not contain enough information to predict both exposure variables, $Q_{x_1}$ will be asymptotically $\chi^2_{L-1}$ distributed where $L$ is the number of SNPs in the estimation. Rejecting the null hypothesis indicates that the SNPs can predict $X_1$ conditional on $X_2$. An equivalent $Q$ statistic for $X_2$ can be calculated by swapping $X_1$ and $X_2$ in equation

2.

The test statistic is calculated for each test statistic individually and the instrument are relevant for the multivariable model if all tests reject the null hypothesis. Generalising to $k$ exposures and using matrix notation, the vector of estimated effects of snp $j$ on each of the exposures other than $x_k$ can be written as: $\hat{\pi}_{-k,j} = (\hat{\pi}_{1,j} \quad \dots \hat{\pi}_{k-1,j} \quad \hat{\pi}_{k+1,j} \quad \dots \quad \hat{\pi}_{K,j})'$ . The relationship between the estimated effect of the SNPs to be tested can be generalised to;

$$\hat{\pi}_{k,j} = \delta \hat{\pi}_{-k,j} + v$$

Where $\delta$ is a $(K-1)$ vector $(\delta_1 \quad \dots \quad \delta_{k-1} \quad \delta_{k+1} \dots \quad \delta_K)$. The Q statistic for $x_k$ can then be written as;

$$Q_{x_k} = \sum_{j=1}^{L} \left( \frac{1}{\sigma^2_{x_k j}} \right) \left( \hat{\pi}_{k,j} - \hat{\delta} \hat{\pi}_{-k,j} \right)^2$$

Where $\hat{\delta}$ is an efficient estimator for $\delta$, and the variance term $\sigma^2_{x_k j}$ is given by;

$$\sigma^2_{x_k,j} = \hat{\delta}^* \Sigma_{V,j} (\hat{\delta}^*)'$$

Where $\hat{\delta}^*$ is the K by 1 vector $(\hat{\delta}_1 \quad \dots \quad \hat{\delta}_{k-1} \quad -1 \quad \hat{\delta}_{k+1} \quad \dots \quad \hat{\delta}_K)$, where $\hat{\delta}_k$ is an efficient estimator for $\delta_k$. $\Sigma_{V,j}$ is the variance covariance matrix for the estimated effect of snp j on each of the exposures, i.e.;

$$\Sigma_{V,j} = \begin{pmatrix} \sigma^2_{1,j} & \sigma_{12,j} & \cdots & \sigma_{1K,j} \\ \sigma_{12,j} & \sigma^2_{2,j} & \cdots & \sigma_{2K,j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K,j} & \sigma_{2K,j} & \cdots & \sigma^2_{K,j} \end{pmatrix}$$

If each $\hat{\pi}_{kj}$ is obtained separately via univariable regressions with an intercept, then the error terms are obtained from the expressions:

$$\sigma^2_{k,j} = \frac{\left(G_j^T G_j\right)^{-1}}{n} \sum_{i=1}^{n} \hat{v}^2_{ki}, \quad \text{and} \quad \sigma_{km,j} = \frac{\left(G_j^T G_j\right)^{-1}}{n} \sum_{i=1}^{n} \hat{v}_{ki} \hat{v}_{mi}$$

$\hat{v}_{kij}$ and $\hat{v}_{mij}$ are the estimated residuals from the $j$'th regression for exposures $k$ and $m$, $k \neq m$.

This $Q$ statistic can then be converted into a F-test for weak instruments by dividing by $L-(k-1)$ where $L$ is the number of SNPs included as instruments and $k$ is the number of exposure variables in the model. This test can then be compared to the critical values tabulated by Stock and Yogo for 1 exposure and $L-(k-1)$ instruments as we are testing the null hypothesis that the instruments explain one fewer exposures than are included in the estimation.

This test statistic requires one piece of data that is not ususally avaliable from GWAS summary statistics $\sigma_{i,j}$ $i \neq j$. Throughout the analysis here we have (unrealistically) assumed this to be zero however, as the proportion of the variance of each exposures explained by any particular SNP is small, this could be estimated

from the phenotypic association between the exposures. Alternatively, for models with only two exposure variables, bounds on the conditional F-statistic can be calculated based on the upper and lower limits that the covariance term could take.

## References


1. Kettunen J, Demirkan A, WÃ¼rtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. Nature communications. 2016
2. Fritsche LG, Igl W, Bailey JNC, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature genetics. 2016, 48(2):134-143
3. Willer C, Schmidt E, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. Nature Genetics. 2013 Nov;45(11):1274-1285.
4. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics. 2018 Oct;50.
5. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. American journal of epidemiology. 2015 Jan 27;181(4):251-60.
6. 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, International Journal of Epidemiology. 2019
7. Bowden J, Del Greco M F, Minelli C, Zhao Q, Lawlor D, Sheehan N, Thompson J, Davey Smith G, Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption, International Journal of Epidemiology. 2019
8. Windmeijer F. Testing Over- and Underidentification in Linear Models, with Applications to Dynamic Panel Data and Asset-Pricing Models. Bristol Economics Discussion Papers 18/696, Department of Economics, University of Bristol, UK, March 2018.


## Code

```
conditionalF <- function(exposures)
{

F.analysis <- data.frame(Fstat[,exposures])
ex.analysis <- data.frame(exp[,exposures])
se.analysis <- data.frame(dat_se[,exposures])

maxF_row <- apply(F.analysis,1,function(x) max(as.numeric(x)))
keep <- as.vector(as.numeric(maxF_row > 10))

F.analysis <- F.analysis*keep
F.analysis[F.analysis == 0] <- NA
F.analysis<- na.omit(F.analysis)
```

```
ex.analysis <- ex.analysis[,1:length(exposures)]*keep
ex.analysis[ex.analysis == 0] <- NA
ex.analysis<- na.omit(ex.analysis)

se.analysis <- se.analysis[,1:length(exposures)]*keep
se.analysis[se.analysis == 0] <- NA
se.analysis<- na.omit(se.analysis)

Q <- data.frame()
F.conditional <- data.frame()
F.mean <- data.frame()

for(j in 1:length(exposures)){

  F.mean <- apply(F.analysis,2,function(b) mean(as.numeric(b)))

  X1 <- ex.analysis[,exposures[j]]
  X2 <- data.frame(ex.analysis[,-which(names(ex.analysis) == exposures[j])])
  X2 <- (matrix(unlist(X2), ncol = length(exposures)-1))

  X1se <- se.analysis[,exposures[j]]
  X2se <- se.analysis[,-which(names(ex.analysis) == exposures[j])]
  X2se <- (matrix(unlist(X2se), ncol = length(exposures)-1))

  delta <- as.vector(lm(X1~ -1 + X2)$coefficients)
  vx <- X1se^2 + (X2se)^2%*%(delta^2)
  Q[1,j] <- sum((1/vx)*((lm(X1~ -1 + X2)$residuals)^2))
  F.conditional[1,j]<-Q[1,j]/length(X1)
  no.snps <- length(X1)

}
F.conditional <- t(F.conditional)
rownames(F.conditional)<-exposures
colnames(F.conditional)<-"Conditional F"

results <- data.frame(F.conditional, F.mean, no.snps)
return(results)


}

Fstrong <- function(exposure)
{
```

```
  F.strong <- data.frame()
  for(j in 1:length(exposure)){

  F.analysis <- data.frame(Fstat[,exposure[j]])
  keep <- as.vector(as.numeric(F.analysis> 10))
  F.analysis <- F.analysis*keep
  F.analysis[F.analysis == 0] <- NA
  F.analysis<- na.omit(F.analysis)

  F.strong[j,1] <- apply(F.analysis,2, mean)
  F.strong[j,2] <- dim(F.analysis)[1]
}

  row.names(F.strong) <- exposure
  colnames(F.strong) <- c("F.stat", "no.snps")
  return(F.strong)

}


rm(list = ls(all=TRUE))

library(data.table)
library(knitr)
library(tidyr)
library(dplyr)
library(devtools)
library(MRChallenge2019)
source("conditionalF.R")
source("Fstrong.R")

dat <- Challenge_dat
dat_se <- data.frame(read.csv("data_incse.txt"))
NMRAdat <- NMRA_dat

names <- NMRAdat$Abbreviation
colnames(dat_se) <- gsub("_", ".", colnames(dat_se))

ids <- as.vector(dat_se$rsid)
row.names(dat_se) <- ids
dat_se <- dat_se[,2:(length(names)+1)]

names <- c("ldl", "hdl", "tg", names)
```

```
exp <- subset(dat, select=c(1,9,12,15,32:149))
pvals <- subset(dat, select=c(11,14,17,150:267))
colnames(exp) <- sub("beta_","",colnames(exp))
names(exp)[names(exp) == 'acAce'] <- 'AcAce'
colnames(pvals) <- sub("p_","",colnames(pvals))

ids <- exp$rsid
row.names(exp) <- ids
row.names(pvals) <- ids

dat_se <- data.frame(dat$se_ldl, dat$se_hdl, dat$se_tg, dat_se)
colnames(dat_se) <- gsub("dat.se_", "", colnames(dat_se))

Fstat <- data.frame()
for(x in 1:length(names)){
for(y in 1:length(ids)){

  Fstat[ids[y],names[x]] <- (exp[ids[y],names[x]]/dat_se[ids[y],names[x]])^2

}
}


F.ind <- Fstrong(names[4:length(names)])
F.ind <- F.ind[order(-F.ind$F.stat),]
topexp <- row.names(F.ind[1:10,])

F.MR <- data.frame(Fstat[,topexp])
ex.MR <- data.frame(exp[,topexp])

maxF_row <- apply(F.MR,1,function(x) max(as.numeric(x)))
keep <- as.vector(as.numeric(maxF_row > 10))

ex.MR <- ex.MR[,1:length(topexp)]*keep
ex.MR[ex.MR == 0] <- NA

MR.all <- (summary(lm(dat$beta_amd~ -1 + ., data = ex.MR,
                      weights = (dat$se_amd)^2)))$coefficients

b<- Fstrong(topexp)
c <- conditionalF(topexp)
```

```
Ftop <- data.frame(c, b)
colnames(Ftop)[4] <- "Ind.F.Stat"
colnames(Ftop)[5] <- "No.snps.Ind"

a <- c("CH2.in.FA", "Bis.DB.ratio", "Bis.FA.ratio")
b <- c( "S.HDL.P",   "S.HDL.L")
c <- c("XL.HDL.TG", "XL.HDL.P", "XL.HDL.PL" ,   "XL.HDL.FC",    "XL.HDL.L" )


FAstrong<- Fstrong(a)
FAcond <- conditionalF(a)
F.FA <- data.frame(FAcond,FAstrong)
S.HDLstrong<- Fstrong(b)
S.HDLcond <- conditionalF(b)
F.S.HDL <- data.frame(S.HDLcond,S.HDLstrong)
XL.HDLstrong<- Fstrong(c)
XL.HDLcond <- conditionalF(c)
F.XL.HDL <- data.frame(XL.HDLcond,XL.HDLstrong)

d <- conditionalF(c("Bis.DB.ratio", "S.HDL.P", "XL.HDL.P"))
e <- Fstrong(c("Bis.DB.ratio", "S.HDL.P", "XL.HDL.P"))
Fsub <- data.frame(d, e)

f <- conditionalF(c("Bis.DB.ratio", "S.HDL.P", "XL.HDL.P", "XL.HDL.TG"))
g <- Fstrong(c("Bis.DB.ratio", "S.HDL.P", "XL.HDL.P", "XL.HDL.TG"))
Fsub2 <- data.frame(f, g)

subexp <- c("Bis.DB.ratio", "S.HDL.P", "XL.HDL.P", "XL.HDL.TG")
F.MR <- data.frame(Fstat[,subexp])
ex.MR <- data.frame(exp[,subexp])
maxF_row <- apply(F.MR,1,function(x) max(as.numeric(x)))
keep <- as.vector(as.numeric(maxF_row > 10))
ex.MR <- ex.MR[,1:length(subexp)]*keep
ex.MR[ex.MR == 0] <- NA

MR.subset <- summary(lm(dat$beta_amd~ -1 + ., data = ex.MR,
                        weights = (dat$se_amd)^-2))$coefficients

weight = (dat$se_amd)^-1 + ((MR.subset[1,1]^2)*(dat_se$Bis.DB.ratio^2))^(-1/2)
+ ((MR.subset[2,1]^2)*(dat_se$S.HDL.P^2))^(-1/2) +
  ((MR.subset[3,1]^2)*(dat_se$XL.HDL.P^2))^(-1/2) +
  ((MR.subset[4,1]^2)*(dat_se$XL.HDL.TG^2))^(-1/2)
MR.subsetup <- summary(lm(dat$beta_amd~ -1 + ., data = ex.MR,
```

```
                          weights = weight))$coefficients
weight = (dat$se_amd)^-1 + ((MR.subsetup[1,1]^2)*(dat_se$Bis.DB.ratio^2))^(-1/2) +
  ((MR.subsetup[2,1]^2)*(dat_se$S.HDL.P^2))^(-1/2) +
  ((MR.subsetup[3,1]^2)*(dat_se$XL.HDL.P^2))^(-1/2) +
  ((MR.subsetup[4,1]^2)*(dat_se$XL.HDL.TG^2))^(-1/2)
MR.subsetup <- summary(lm(dat$beta_amd~ -1 + ., data = ex.MR,
                          weights = weight))$coefficients
weight = (dat$se_amd)^-1 + ((MR.subsetup[1,1]^2)*(dat_se$Bis.DB.ratio^2))^(-1/2) +
  ((MR.subsetup[2,1]^2)*(dat_se$S.HDL.P^2))^(-1/2) +
  ((MR.subsetup[3,1]^2)*(dat_se$XL.HDL.P^2))^(-1/2) +
  ((MR.subsetup[4,1]^2)*(dat_se$XL.HDL.TG^2))^(-1/2)

MR.subsetup <- summary(lm(dat$beta_amd~ -1 + ., data = ex.MR, weights = weight))$coefficients



exposures <-c("ldl", "hdl", "tg")
a<- Fstrong(exposures)
b <- conditionalF(exposures)
Fchol <- data.frame(b, a)
HDL <- dat$beta_hdl
LDL <- dat$beta_ldl
TG <- dat$beta_tg

glc.exp <- c("hdl", "ldl", "tg")
F.MR <- data.frame(Fstat[,glc.exp])
ex.MR <- data.frame(exp[,glc.exp])
maxF_row <- apply(F.MR,1,function(x) max(as.numeric(x)))
keep <- as.vector(as.numeric(maxF_row > 10))
ex.MR <- ex.MR[,1:length(glc.exp)]*keep
ex.MR[ex.MR == 0] <- NA



MR.results <- (summary(lm(dat$beta_t2d ~ -1 + ., data = ex.MR,
                          weight = (dat$se_t2d)^-1)))$coefficients

weight = (dat$se_t2d)^-1 + ((MR.results[1,1]^2)*(dat$se_hdl^2))^(-1/2) +
  ((MR.results[2,1]^2)*(dat$se_ldl^2))^(-1/2) +
    ((MR.results[3,1]^2)*(dat$se_tg^2))^(-1/2)

MR.resultsup <- (summary(lm(dat$beta_t2d ~ -1 + ., data=ex.MR,
                            weight = weight)))$coefficients
```