MR Data Challenge 2019 — The role of lipoprotein subfractions in coronary artery disease

Qingyuan Zhao (on behalf of the Penn team)

June 5th, 2019

1 Participants

Qingyuan Zhao[1], Jingshu Wang[1], Zhen Miao[2], Nancy Zhang[1], Sean Hennessy[2], Dylan Small[1], Dan Rader[2]

[1] Department of Statistics, Wharton School, University of Pennsylvania

[2] Perelman School of Medicine, University of Pennsylvania

2 Motivation

Earlier MR studies (Voight et al. 2012; Zhao, Chen, et al. 2019) found evidence of heterogeneity for the effect of HDL cholesterol on coronary artery disease (CAD). In this report, the research question we aim to explore is:

Do lipoprotein subfractions have heterogeneous effects on CAD?

To answer this question, we will conduct a MR screening study that estimates the causal effect of each subfraction on CAD. Some highlighting features of our analysis include:

- **Genome-wide three-sample design**: This design allows us to utilize insturments that are only weakly correlated with the subfraction trait (may not be genome-wide significant). This is crucial because usually only a handful of genetic variants have genome-wide significant association with subfraction traits.
- **State-of-the-art statistical method**: We used Robust Adjusted Profile Score (RAPS) which is not biased by individually weak instruments, as long as the overall instrument strength is not weak. RAPS is also robust to balanced and/or sparse pleiotropy; in particular, RAPS is asymptotically unbiased if the InSIDE (INstrument Strength Independent of Direct Effect) assumption is satisfied.
- **Multivariate MR**: We applied a novel extension of RAPS to multivariate exposures. This allows us to assess whether the (potential) effect of any lipoprotein subfraction is independent of the major lipid traits (HDL-C, LDL-C, TG).

3 Data

To maximize the statistical power, we did not use dataset provided by the challenge organizers. Instead, we generated our own data frames from raw summary datasets as listed in the table below. For the lipoprotein subfractions, besides the Kettunen et al. (2016) dataset, we used the summary results of another GWAS by Davis et al. (2017). This can help us to select specific instruments for each subfraction trait. This also allows us to obtain two MR estimates for each subfraction exposure using different datasets.

Phenotype	1st Author	PubMed ID	URL to summary dataset
Traditional lipids	Hoffman	29507422	https://www.ebi.ac.uk/gwas/studies/GCST007141
Traditional lipids	Willer	24097068	http://csg.sph.umich.edu/abecasis/public/lipids2013/
Subfractions	Davis	29084231	http://csg.sph.umich.edu/boehnke/public/metsim-2017-lipoproteins/
Subfractions	Kettunen	27005778	http://www.computationalmedicine.fi/data#NMR_GWAS
Heart attack	Abbott	Interim	http://www.nealelab.is/uk-biobank/
CAD	Nikpay	26343387	http://www.cardiogramplusc4d.org/data-downloads/
CAD	Nelson	28714975	http://www.cardiogramplusc4d.org/data-downloads/

See References at the end of this report for detailed citations of the GWAS datasets we used: Hoffmann et al. (2018), Willer et al. (2013), Davis et al. (2017), Kettunen et al. (2016), Abbott et al. (2018), Nikpay et al. (2015), Nelson et al. (2017).

We further preprocessed the data (see the Software section below) using the three-sample summary-data design in Zhao, Chen, et al. (2019). We obtained three datasets for the MR analysis as descrbied in the table below. Briefly speaking, the selection GWAS is used to select independent genetic instruments (distance $\geq$ 10 megabase pairs, linkage disequilibrium $r^2 \leq 0.001$) that are associated with the exposure(s) under investigation. The MR dataset is then created by obtaining the summary statistics in the exposure and outcome GWAS corresponding to the selected instruments. Two of our study designs are univariate, that is only the lipoprotein subfraction is used as the exposure. To evaluate whether any discovered subfraction effect is independent of the traditional lipid factors, we also conducted a multivariate MR study where the lipoprotein subfraction and the "other" traditional lipid factors are used as exposures. For example,

- For s_hdl_p, the multivariate MR analysis also adjusts for LDL-C and TG.
- For s_ldl_p, the multivariate MR analysis also adjusts for HDL-C and TG.
- For s_vldl_p, the multivariate MR analysis also adjusts for HDL-C and LDL-C.

Name	Selection	Exposure	Outcome
Univariate I	Hoffman (GERA)	Davis (METSIM)	Nikpay (CARDIoGRAM)
Univariate II	Davis (METSIM)	Kettunen	Abbott (UKBB)
Multivariate	Hoffman (GERA) + Davis (METSIM)	Willer (GLGC) + Kettunen	Nelson (CARDIoGRAM & UKBB)

Notice that in all the designs, the selection GWAS has no overlapping sample with the exposure and outcome GWAS. This makes sure that our statistical analysis does not suffer from selection bias (winner's curse).

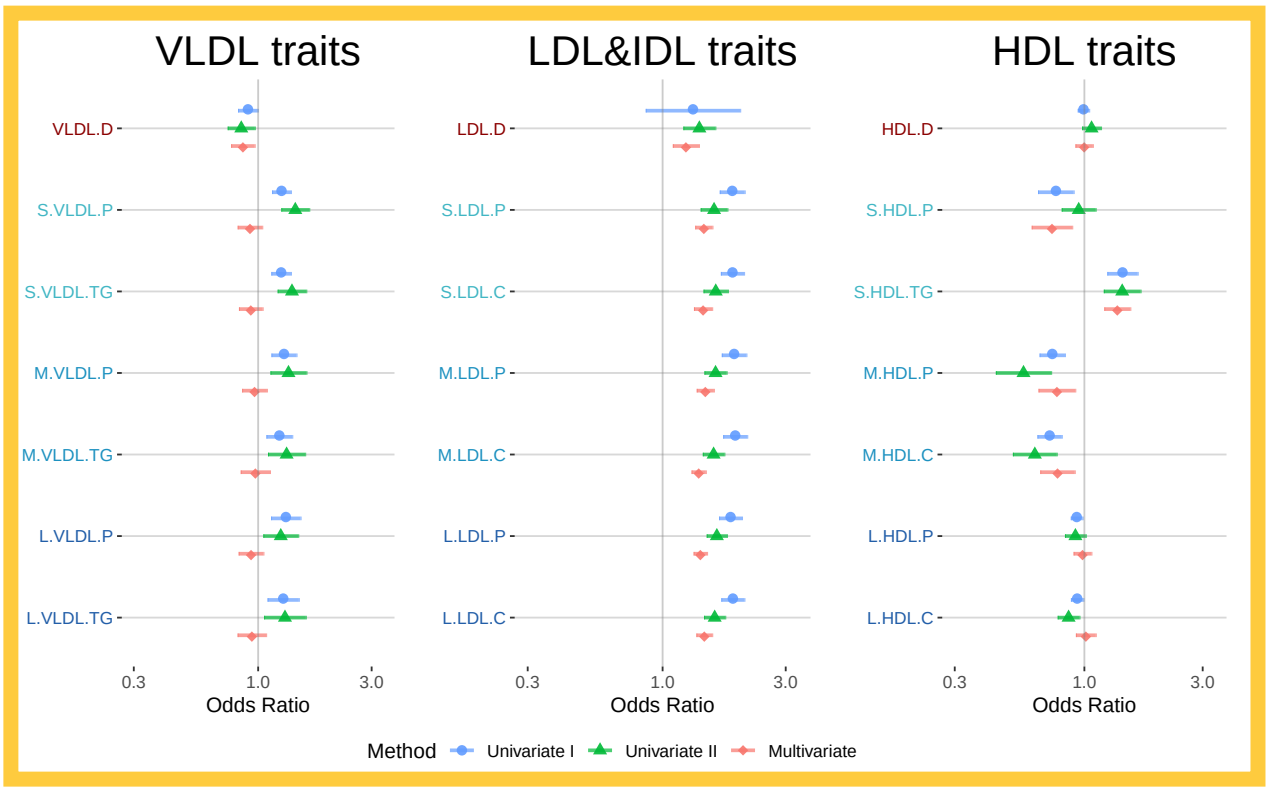

Figure 1: Main results.

When selecting the instruments, we did not put restriction on how strong the instrument (aka p-value in the selection GWAS) need to be. This is because the RAPS method can automatically put optimal weight on the instruments according to their strength (Zhao, Chen, et al. 2019). This feature has not been implemented for GRAPPLE (multivariate extension to RAPS), so we require the instruments used in the multivariate MR analysis to be associated with at least one exposure with p-value $\leq 10^{-4}$ in the selection GWAS.

4 Analysis methods

For univariate MR, we used robust adjusted profile score (RAPS) (Zhao, Wang, et al. 2019; Zhao, Chen, et al. 2019). For multivariate MR, we used an extension of RAPS called GRAPPLE (Wang et al, 2019, forthcoming). Both methods are implemented in the mr.raps R package. We further compared with other univariate MR methods including inverse variance weighting (IVW), weighted median, and Egger regression in the Software section.

5 Results

The main results are shown in Figure 1. A further comparison of RAPS with other univariate MR methods can be found in Figure 2 in the Software section. From these plots we can make the following observations:

1. Genetically-determined LDL and VLDL subfractions had (mostly) homogeneous associations with CAD

in univariate MR.

2. However, **genetically-determined HDL subfractions had heterogeneous associations with CAD** in univariate and multivariate MR.
3. In particular, genetically-determined medium HDL traits had significantly negative associations with CAD in univariate and multivariate MR (that adjusts for LDL-C and TG).
4. Genetically-determined triglycerides in small HDL had significantly positive associations with CAD
5. Using weak instruments substantially increases the statistical power of RAPS and does not appear to bias its point estimate. All the other univariate MR methods are biased by weak instruments (Figure 2)
6. The diagnostic plots for m_hdl_p (concentration of medium HDL particles) (Figure 3) do not suggest evidence against the InSIDE (INstrument Strength Independent of Direct Effect) assumption. This is the crucial assumption so that the results of RAPS may have a causal interpretation.

In conclusion, we find strong evidence that HDL subfractions may have heterogeneous effects on CAD. The relationship of this finding with the HDL function hypothesis (Rader and Hovingh 2014) requires further investigations.

6 Software

6.1 Setting up the environment

Install the necessary software packages:

```
devtools::install_github("qingyuanzhao/mr.raps", ref = "multivariate")
devtools::install_github("qingyuanzhao/TwoSampleMR")
# this contains a small modification to the MRCIEU master repo which changes
# the default option for mr_raps
```

Load the software packages:

```
library(mr.raps)
library(TwoSampleMR)
```

Our preprocessing relies on local installation of PLINK and reference panel:

```
plink_exe <- paste0("plink")
plink_refdat <- paste0("ld_files/data_maf0.01_rs")
```

This section will contain a illustration of our analysis using some selected HDL subfractions:

```
traits <- c("s_hdl_p", "s_hdl_tg", "m_hdl_p", "m_hdl_c", "l_hdl_p", "l_hdl_c")
```

All the datasets used in this analysis is stored in the local "data" folder.

```
list.files("data/")
```

```
##  [1] "cardiogramplusc4d_ukbb_cad.rda" "davis_l_hdl_c.rda"
##  [3] "davis_l_hdl_p.rda"              "davis_m_hdl_c.rda"
##  [5] "davis_m_hdl_p.rda"              "davis_s_hdl_p.rda"
```

```
##  [7] "davis_s_hdl_tg.rda"            "gera_hdl.rda"
##  [9] "gera_ldl.rda"                  "gera_tg.rda"
## [11] "kettunen_l_hdl_c.rda"          "kettunen_l_hdl_p.rda"
## [13] "kettunen_m_hdl_c.rda"          "kettunen_m_hdl_p.rda"
## [15] "kettunen_s_hdl_p.rda"          "kettunen_s_hdl_tg.rda"
## [17] "mi_ukbb.rda"                   "willer_hdl.rda"
## [19] "willer_ldl.rda"                "willer_tg.rda"
```

They are structured data frames (saved in the "dat" object) that can be loaded in R. For example,

```
load("data/gera_hdl.rda")
head(dat, 3)
```

```
##         SNP Chromosome Position effect_allele   eaf samplesize     beta
## 1 rs3094315          1   752566             G 0.164      94235 -0.00505
## 2 rs2905035          1   775659             A 0.144      83482  0.00529
## 3 rs2980319          1   777122             A 0.267      10753  0.00371
##   pval other_allele          se
## 1 0.32            A 0.005078144
## 2 0.80            G 0.020880444
## 3 0.76            T 0.012144790
```

6.2 Data preprocessing for univariate MR:

As described in the Data section, we preprocess the raw GWAS summary data using the `getInput` function in the `mr.raps` package. At this point, we do not set a instrument strength threshold for selecting the instruments.

```
data <- list()
for (trait in traits) {
  data[[trait]] <- getInput(sel.files = paste0("data/davis_", trait, ".rda"),
                            exp.files = paste0("data/kettunen_", trait, ".rda"),
                            out.file = "data/mi_ukbb.rda",
                            plink_exe, plink_refdat, p.thres = 1)
}
for (trait in traits) {
  data[[trait]]$id.exposure <- NA
  data[[trait]]$id.outcome <- NA
  data[[trait]]$exposure <- trait
  data[[trait]]$outcome <- "MI"
  data[[trait]]$mr_keep <- TRUE
}
save(data, file = "univariate_data.rda")
```

6.3 Data preprocessing for multivariate MR

We preprocess the data for multivariate MR in the same way. Because empirical Bayes shrinkage has not been implemented in GRAPPLE yet, we only include instruments whose minimum p-value in the selection GWAS is smaller than $10^{-}4$. The `calCor` function estimates the sample overlap between the exposure and outcome GWAS.

```
data <- list()
corr <- list()
for (trait in traits) {
    sel.files <- c(paste0("data/gera_", c("hdl", "ldl", "tg"), ".rda"),
                   paste0("data/davis_", trait, ".rda"))
    exp.files <- c(paste0("data/willer_", c("ldl", "tg"), ".rda"),
                   paste0("data/kettunen_", trait, ".rda"))
    out.file <- "data/cardiogramplusc4d_ukbb_cad.rda"
    data[[trait]] <- getInput(sel.files, exp.files, out.file,
                              plink_exe, plink_refdat, p.thres = 1e-4)
    corr[[trait]] <- calCor(sel.files, exp.files, out.file,
                plink_exe, plink_refdat)
}
save(data, corr, file = "multivariate_data.rda")
```

6.4 Results for univariate MR

The next code chunk applies four univariate MR methods (Egger, weighted median, IVW, RAPS) to the dataset we just created. To compare the performance of these methods with weak instruments, we consider two thresholds for instrument selection: 10^{-6} and 1. Finally, the p-values are adjusted by Bonferroni's correction.

```
load("univariate_data.rda")

methods <- c("mr_egger_regression", "mr_weighted_median", "mr_ivw", "mr_raps")
parameters <- default_parameters()
parameters$shrinkage <- TRUE

out.all <- data.frame()
for (trait in traits) {
  for (pval.sel in c(1, 1e-6)){
    out <- mr(subset(data[[trait]], pval.selection < pval.sel),
              parameters = parameters,
              method_list = methods)
    out$pval.sel <- pval.sel
    out.all <- rbind(out.all, out)
  }
```

```
}
out.all$pval.adjusted <- p.adjust(out.all$pval, "bonferroni")
levels(out.all$method) <- c("IVW", "Egger", "RAPS", "W. Median")
```

The results of the univaraite MR analysis are shown in Figure 2, which is generated by the next code chunk. The 95% confidence intervals that are still significant after Bonferroni's adjustment are shown in blue.

```
library(ggplot2)
ggplot(out.all) + aes(x = method, y = b, ymin = b - 1.96 * se, ymax = b + 1.96 * se,
                      col = pval.adjusted < 0.05) +
  geom_point() + geom_errorbar() + geom_hline(yintercept = 0, linetype = "dashed") +
  facet_grid(exposure ~ pval.sel) + coord_flip() +
  theme_light(base_size = 18) + theme(legend.position = "bottom")
```

The `mr.raps` package further implements diagnostics and falsification test of the MR analysis. Below is an example for m_hdl_p (concentration of medium HDL particles). In Figure 3, the left plot shows that the standardized residual (y-axis) is roughly mean 0 and independent of the instrument weight (x-ais); the right plot shows that the empirical distribution of the standardized residuals is very close to a standard normal distribution. Both of them are necessary consequences of the modeling assumptions of RAPS (Zhao, Chen, et al. 2019), so these diagnostic plots look "normal" and indicate no evidence against our modeling assumptions.

```
mr.raps(data[["m_hdl_p"]], shrinkage = TRUE)
```

```
## Estimated causal effect: -0.0153, standard error: 0.00311, p-value: 8.84e-07.
## Estimated overdispersion variance: 1.67e-08, standard error: 8.61e-09, p-value: 0.052.
## ANOVA test: are the weights and residuals independent?
## Analysis of Variance Table
##
## Response: std.resids
##                                                            Df  Sum Sq Mean Sq
## bs(weights, knots = quantile(weights, 1:df/(df + 1)))   89   92.97 1.04463
## Residuals                                                1637 1603.73 0.97968
##                                                            F value Pr(>F)
## bs(weights, knots = quantile(weights, 1:df/(df + 1)))  1.0663 0.3203
## Residuals

## $beta.hat
## [1] -0.01527959
##
## $tau2.hat
## [1] 1.673251e-08
##
## $beta.se
## [1] 0.00310823
##
## $tau2.se
```

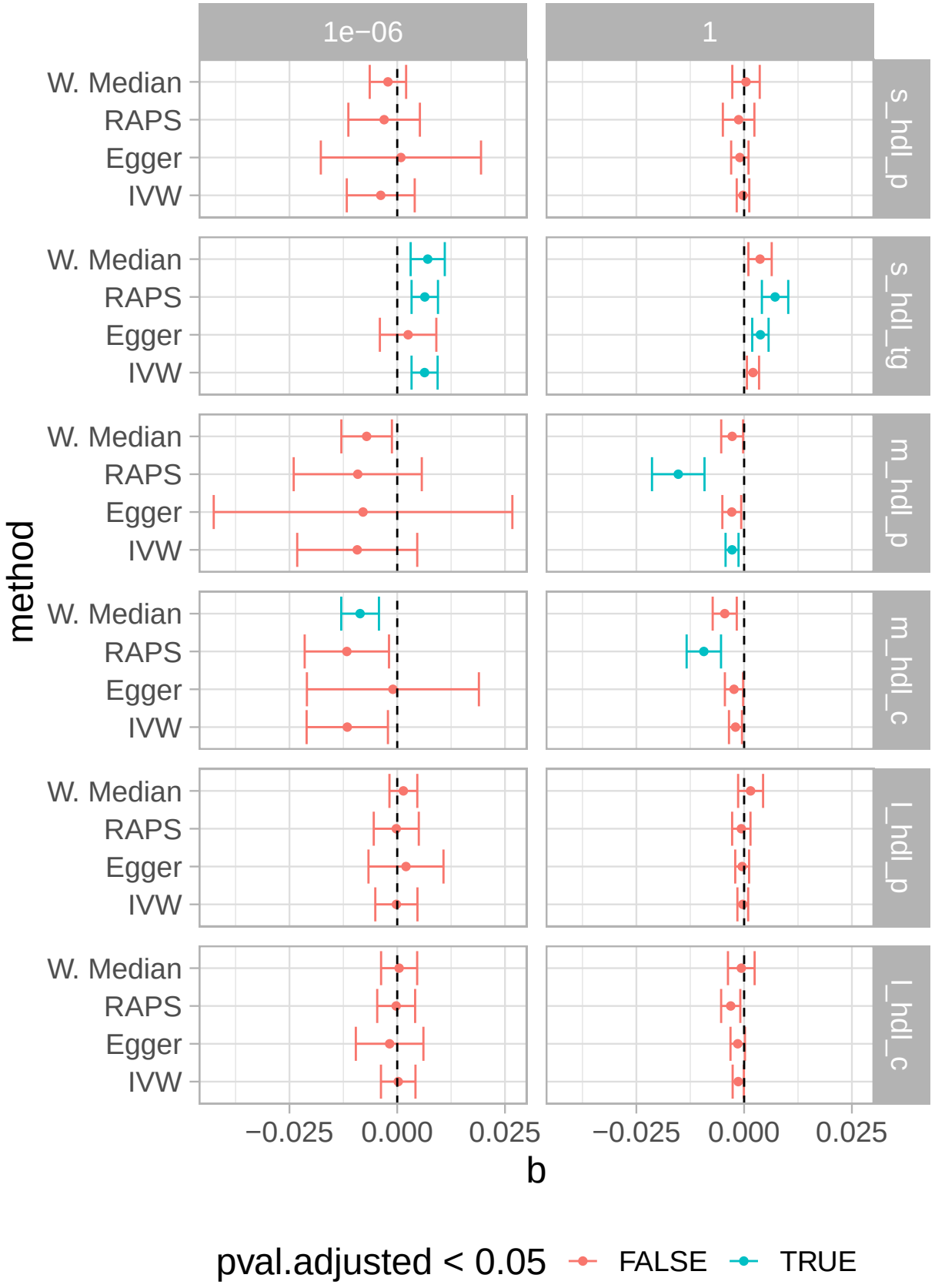

Figure 2: Results for selected HDL subfractions.

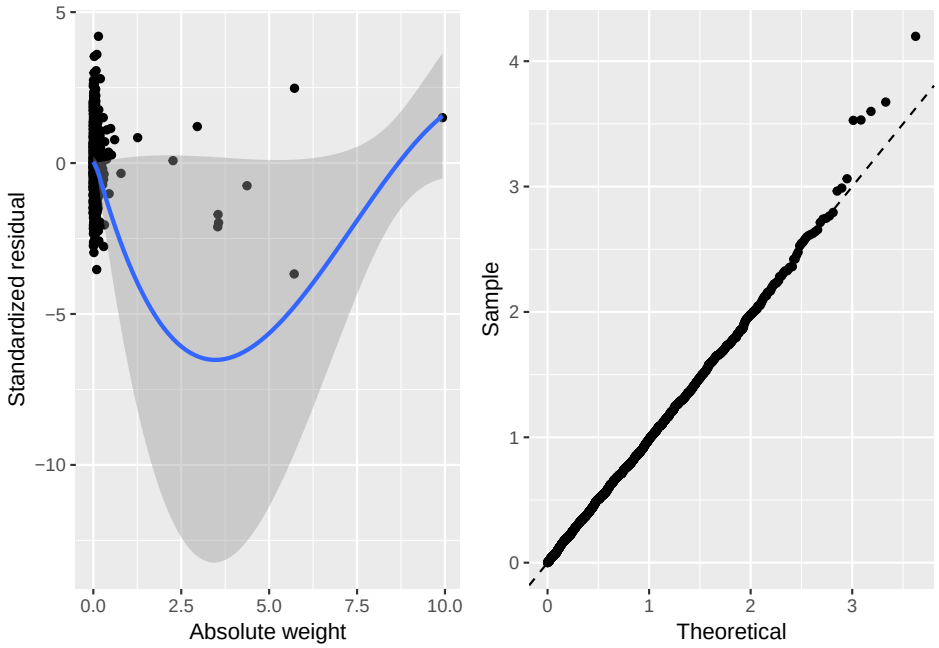

Figure 3: Diagnostic plots for RAPS.

```
## [1] 8.611568e-09
```

6.5 Results for multivariate MR

The next code chunk applies GRAPPLE (multivariate extension to RAPS) to the multivariate MR dataset we created. The first three exposures in the multivariate MR are HDL-C, LDL-C, and TG. The fourth and last exposure is a lipoprotein subfraction trait.

```
load("multivariate_data.rda")
result <- list()
for (trait in traits) {
    result[[trait]] <- grappleRobustEst(data[[trait]]$beta_exp,
                                        data[[trait]]$data_out$beta,
                                        data[[trait]]$se_exp,
                                        data[[trait]]$data_out$se,
                                        cor.mat = corr[[trait]],
                                        loss.function = "huber")
}
```

Here are the point estimates, standard errors, and two-sided p-values of our multivariate MR analysis:

```
print(b <- sapply(result, function(x) x$beta.hat[3]))
```

```
##     s_hdl_p    s_hdl_tg     m_hdl_p     m_hdl_c     l_hdl_p     l_hdl_c
## -0.30100780  0.30576565 -0.25522155 -0.24962025 -0.01746443  0.01361051
```

```r
print(se <- sapply(result, function(x) sqrt(x$beta.var[3, 3])))
```

```
##    s_hdl_p   s_hdl_tg    m_hdl_p    m_hdl_c    l_hdl_p    l_hdl_c
## 0.09602275 0.06175015 0.08713211 0.08181847 0.04225070 0.04715766
```

```r
print(pval <- 2 * pnorm(-abs(b)/se))
```

```
##      s_hdl_p     s_hdl_tg      m_hdl_p      m_hdl_c      l_hdl_p
## 1.719976e-03 7.358362e-07 3.399086e-03 2.281541e-03 6.793484e-01
##      l_hdl_c
## 7.728743e-01
```

References

Abbott, Liam, Sam Bryant, Claire Churchhouse, Andrea Ganna, Daniel Howrigan, Duncan Palmer, Ben Neale, Raymond Walters, Caitlin Carey, and The Hail team. 2018. "Round 2 GWAS Results of Thousands of Phenotype in the UK BioBank." http://www.nealelab.is/uk-biobank/.

Davis, James P, Jeroen R Huyghe, Adam E Locke, Anne U Jackson, Xueling Sim, Heather M Stringham, Tanya M Teslovich, et al. 2017. "Common, Low-Frequency, and Rare Genetic Variants Associated with Lipoprotein Subclasses and Triglyceride Measures in Finnish Men from the Metsim Study." *PLoS Genetics* 13 (10): e1007079.

Hoffmann, Thomas J, Elizabeth Theusch, Tanushree Haldar, Dilrini K Ranatunga, Eric Jorgenson, Marisa W Medina, Mark N Kvale, et al. 2018. "A Large Electronic-Health-Record-Based Genome-Wide Study of Serum Lipids." *Nature Genetics* 50 (3): 401.

Kettunen, Johannes, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, et al. 2016. "Genome-Wide Study for Circulating Metabolites Identifies 62 Loci and Reveals Novel Systemic Effects of Lpa." *Nature Communications* 7: 11122.

Nelson, Christopher P, Anuj Goel, Adam S Butterworth, Stavroula Kanoni, Tom R Webb, Eirini Marouli, Lingyao Zeng, et al. 2017. "Association Analyses Based on False Discovery Rate Implicate New Loci for Coronary Artery Disease." *Nature Genetics* 49 (9): 1385.

Nikpay, Majid, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, et al. 2015. "A Comprehensive 1000 Genomes–Based Genome-Wide Association Meta-Analysis of Coronary Artery Disease." *Nature Genetics* 47 (10): 1121.

Rader, Daniel J, and G Kees Hovingh. 2014. "HDL and Cardiovascular Disease." *Lancet* 384 (9943): 618–25.

Voight, Benjamin F, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, et al. 2012. "Plasma HDL Cholesterol and Risk of Myocardial Infarction: A Mendelian Randomisation Study." *Lancet* 380 (9841): 572–80.

Willer, Cristen J, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, et al. 2013. "Discovery and Refinement of Loci Associated with Lipid Levels." *Nature Genetics* 45 (11): 1274.

Zhao, Qingyuan, Yang Chen, Jingshu Wang, and Dylan S Small. 2019. "Powerful Genome-Wide Design and Robust Statistical Inference in Two-Sample Summary-Data Mendelian Randomization." *International Journal of Epidemiology* to appear.

Zhao, Qingyuan, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small. 2019. "Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score." *Annals of Statistics* to appear.