BIOINFORMATICS
MEDIAN RANDOMIZATION DATA REPORT

Pei-Chun, Chen

Department of Statistics, National Cheng Kung University
r26121073@gs.ncku.edu.tw

1. INTRODUCTION

The ”TWBrs” dataset includes information on 13,899 samples, consisting of 13,899 female and 0 male participants, along with 10,000 SNPs. And the complemented file named ”phenotype_TWBrs.csv” which outlines the characteristics of the dataset.

My goal is to perform a Genome-Wide Association Study (GWAS) to findout SNPs that show significant associations with the specified exposure. These identified SNPs will serve as instrumental variables for investigating potential causal effects. Lastly, I will visualize the analysis results to enhance their interpretability.

2. QUALITY CONTROL

I removed samples or loci with more than 2% missing values and deleted variants with a minor allele frequency (MAF) below 5%. Additionally, I excluded variants that did not conform to Hardy-Weinberg equilibrium, using a significance threshold of 10^{6} for hypothesis testing. After these filters, I was left with 9,296 variants (SNPs). The relevant details are summarized in table 1.

2.1. Quality Control ”Code” in plink

```
plink --bfile TWBrs --geno 0.02 --make-bed --
    out TWBrs_geno
plink --bfile TWBrs_geno --mind 0.02 --make-
    bed --out TWBrs_mind
plink --bfile TWBrs_mind --maf 0.05 --make-
    bed --out TWBrs_maf
plink --bfile TWBrs_maf --hwe 1e-6 --make-bed
     --out TWBrs_hardy
```

3. MODELING

Next, I want to investigate the causal relationship between Drinking and Smoking. Therefore, I conduct modeling for both traits, ”Drinks” and ”Smokes”, as they are binary outcomes. I utilize logistic regression for fitting the models to these traits.

3.1. Modeling ”Code” in plink

```
plink --bfile TWBrs_hardy --logistic --ci
    0.95 --pheno phenotype_TWBrs.csv --mpheno
     1 --out DrinksRegression
plink --bfile TWBrs_hardy --logistic --ci
    0.95 --pheno phenotype_TWBrs.csv --mpheno
     2 --out SmokeRegression
```

4. GENOME-WIDE ASSOCIATION STUDIES (GWAS)

I performed a GWAS analysis on SNPs associated with ”Drink”, setting the threshold for multiple comparisons at 5×10^{-4}. The rationale behind this is to relax the criteria in order to identify a larger number of SNPs that are significantly associated with the traits. Four significant SNPs associated with ”Drink” were identified. Manhattan plots were generated to visualize these associations. See Figure 1.

4.1. Manhattan Plot

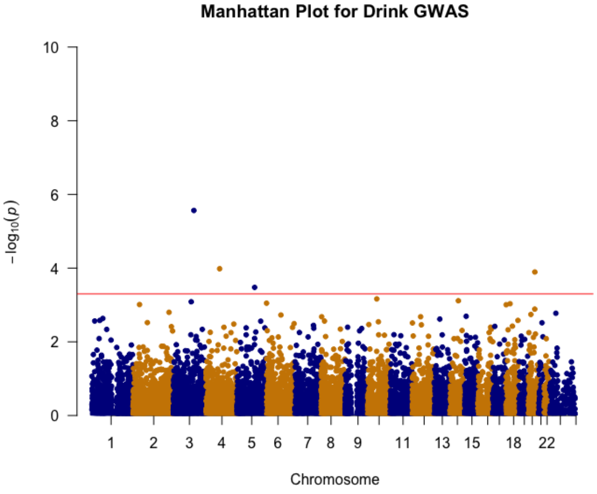


Fig. 1: Manhattan Plot for Drink

Table 1: Testing Data Performance of Segmentation

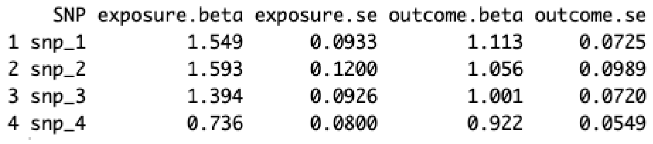
	Threshold	Variants Removed	Variants Remaining
missingness per marker	– geno 0.02	525	9475
missingness per individual	– mind 0.02	0	9475
minor allele frequency	–maf 0.05	110	9365
Hardy-Weinberg equilibrium	–hwe e-6	69	9296

5. MENDELIAN RANDOMIZATION

In this section, I use the four SNPs significantly associated with trait ”drink”, as identified in the previous GWAS analysis, as instrumental variables. Alcohol consumption, herein referred to as ”drink”, is considered as the exposure variable, with the objective of elucidating the causal relationship with the outcome of interest: smoking behavior.

The motivation for examining this relationship arises from the prevalence of alcohol use in typical workplace gatherings or at entertainment venues. I am interested in whether the propensity to drink in these contexts leads to smoking, with the intention to dissect this behavior by examining genetic differences.

5.1. MR_input ”Report” in R



```
    SNP exposure.beta exposure.se outcome.beta outcome.se
1 snp_1          1.549      0.0933        1.113     0.0725
2 snp_2          1.593      0.1200        1.056     0.0989
3 snp_3          1.394      0.0926        1.001     0.0720
4 snp_4          0.736      0.0800        0.922     0.0549
```

Fig. 2: MR_Input

5.2. MR methods Report

Referencing the paper [1], I applied four Mendelian Randomization methods: IVW (Inverse Variance Weighted), MR-Egger, Weighted Median, and Weighted Mode. The results are reported below. It can be observed that the causal estimators from all four methods are significantly different from zero, indicating consistency.

```
------------------------------------------------------------------
Method  Estimate  Std  Error  95%  CI          p-value
   IVW      0.783       0.111  0.565,  1.000     0.000
------------------------------------------------------------------
       Method  Estimate  Std  Error  95%  CI          p-value
   MR-Egger       0.186       0.092  0.005,  0.366     0.044
(intercept)       0.781       0.115  0.556,  1.007     0.000
------------------------------------------------------------------
                  Method  Estimate  Std  Error  95%  CI          p-value
Weighted  median  method       0.718       0.045  0.629,  0.807     0.000
------------------------------------------------------------------
Method  Estimate  Std  Error  95%  CI          p-value
   MBE      0.715       0.045  0.626,  0.804     0.000
------------------------------------------------------------------
```

5.3. MR_forest Plot

In addition to the previous methods, I incorporated three additional methods: Simple Median, Maximum Likelihood, and Contamination Mixture. I also generated their forest plots, which demonstrate that the 95% confidence intervals do not include zero. This indicates that there is 95% confidence that the causal estimates are significantly non-zero, as shown in Figure 3.

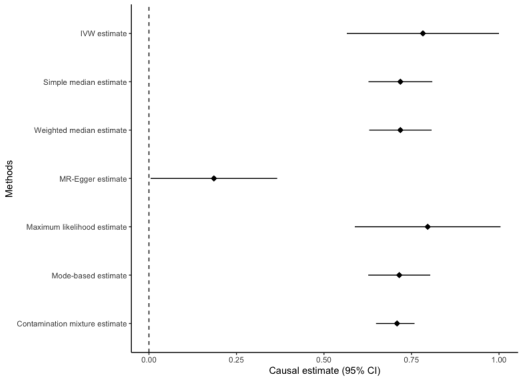


Fig. 3: MR_forest Plot

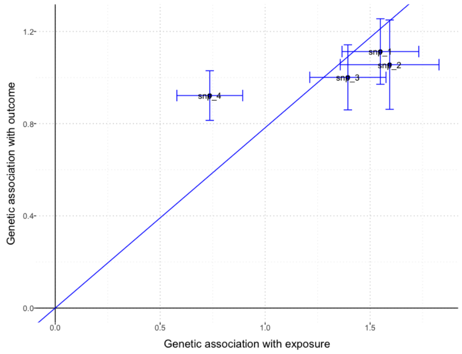


Fig. 4: Genetic Association Plot

5.4. Genetic association with exposure and outcome

The figure above illustrates the genetic association with exposure and outcome. It can be observed that when the genetic association of my SNPs with the exposure is positive, the genetic association with the outcome is also positive. Therefore, overall, the genetic association with exposure and outcome exhibits a consistent direction. See Figure 4.

6. CONCLUSIONS

In this analysis, I introduced four SNPs from a previous GWAS as my instrumental variables and employed various Mendelian randomization methods to infer the causal effect of ”drink” on ”smoke.” The results of the analysis indicate that, under the condition that the instrumental variables, namely the SNPs, are fixed and the three basic assumptions of MR are satisfied, there is a significant positive causal relationship between the trait ”drink” and ”smoke.”

Therefore, the conclusion drawn from my study is that drinking indeed influences smoking, establishing a positive relationship between the two.

7. REFERENCES

[1] Clarke T. K. Adams M. J. McIntosh A. M. Davey Smith G. Jung J. Lohoff F. W. Rosoff, D. B. Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample mendelian randomization study with 780,000 participants. *Molecular psychiatry*, 26(4):1119–1132, 2021.