PH240C Final Project

Causal Impact of Smoking on Coronary Artery Disease: A Mendelian Randomization Analysis

Riddhi Sera

2023-12-15

Abstract

This project explores the causal relationship between smoking and Coronary Artery Disease (CAD) using Mendelian Randomization (MR), a method that leverages genetic variants as instrumental variables to infer causality from observational data. Motivated by the need to understand whether the observed associations between smoking and CAD in epidemiological studies reflect a true causal relationship, the project employs MR as a robust alternative to randomized controlled trials, which are not feasible in this context. The analysis utilized the TwoSampleMR package in R, implementing a multi-faceted approach with methods including Inverse Variance Weighted (IVW), MR Egger Regression, and the Weighted Median Approach. Each method offered unique insights while addressing potential biases such as pleiotropy and invalid instrumental variables. The results consistently indicated a significant causal effect of smoking on CAD across all methods, albeit with variations in the magnitude of the effect. This convergence, despite methodological differences, strengthens the inference of a genuine causal link, suggesting a critical need for public health initiatives to address smoking as a key modifiable risk factor for CAD.

1 Introduction

1. In this project, I have utilized data from the Genome Wide Association Studies Catalog (GWAS) [https://www.ebi.ac.uk/gwas/home] library for my analysis. Specifically, I conducted searches for genetic variants associated with smoking and Coronary Artery Disease (CAD) to compile a comprehensive list of SNPs from the database.

   GWAS summary statistics were obtained from the GWAS Catalog for both smoking (study ID: GCST007327) and CAD (study ID: GCST005194). The choice of these particular datasets was driven by their large dataset size and high association counts, which indicate a strong genetic link to the traits of interest. Moreover, to mitigate potential confounding factors, I ensured that the ethnicity of all the samples was consistent, specifically of European descent. This careful selection criteria aims to enhance the validity of the Mendelian Randomization analysis by reducing the influence of population stratification.

2. This project aims to identify the causal effect of smoking on Coronary Artery Disease (CAD) through

Mendelian Randomization (MR). The link between smoking and CAD has been extensively studied, with a wealth of literature underscoring its negative impact. However, these claims predominantly arise from observational studies, which, while indicative, do not establish causation definitively. Only randomized controlled trials (RCTs) can conclusively test causation, but such trials on smoking are neither ethical nor feasible.

The relationship between lifestyle factors and heart disease has been a significant focus in epidemiological research. For instance, a plethora of case control and cohort studies have suggested that vitamin E is inversely associated with the risk of CHD (5,6). However, this was contradicted by a meta-analysis of 7 RCTs, which showed no significant effect of vitamin E on cardiovascular events (odds ratio, 0.98; 95% CI, 0.94 to 1.03) (7). This discrepancy highlights the limitations of observational studies in establishing causality.

Similar concerns apply to the relationship between smoking and CAD. While the harmful effects of smoking are well-documented, the extent to which smoking is causally linked to CAD needs further exploration. Studies like (8) & (9) have shown a strong association between smoking and increased risk of CAD. However, these are observational in nature.

The Mendelian Randomization approach used in this project offers a solution to this issue. This method uses genetic variants as instrumental variables to infer causal relationships, providing a more robust assessment than traditional observational studies. The genetic component in smoking behavior and its relationship to CAD has been explored in studies like Klarin et al. (10), but not extensively in the context of Mendelian Randomization.

This project, therefore, seeks to bridge the gap between correlation and causation in the context of CAD and smoking. Our investigation is grounded in the hypothesis that genetic predispositions to smoking behavior are causally linked to CAD outcomes. The insights gained could have significant implications for public health policies and interventions, providing a deeper understanding of the true causal impact of smoking on CAD.

3. The statistical methodology I am currently employing for this research is Mendelian Randomization (MR). This approach utilizes genetic variants, specifically single nucleotide polymorphisms (SNPs), as instrumental variables to infer causality between an exposure (in this case, smoking) and an outcome (Coronary Artery Disease, CAD). The rationale behind using MR lies in the genetic variants' association with the exposure, which is assumed to be random with respect to confounders that typically plague observational studies.

   Mendelian Randomization has emerged as a powerful tool in epidemiological studies to address the issue of causality, a concept often discussed but seldom proved in observational research. The method's theoretical underpinning is grounded in Mendel's laws of inheritance, suggesting that alleles segregate independently of environmental factors that may confound traditional epidemiological analyses (5).

   A significant advantage of MR is its ability to mitigate reverse causation and confounding, two major limitations of observational studies, by using genetic variants as proxies for modifiable risk factors (Lawlor et al., 2008). Several studies have successfully utilized MR to explore the causal effects of risk factors on disease outcomes. For example, MR has been used to investigate the relationship between body mass index and various health outcomes, providing insights that have guided public

health interventions (Nordestgaard et al., 2012).

In the context of smoking and CAD, MR is particularly suited because randomized controlled trials (RCTs) involving smoking are neither ethical nor practical. Davey Smith and Hemani (5) highlight the use of MR to ascertain the causal role of smoking in diseases, where the authors leveraged genetic instruments that affect smoking behavior to understand its impact on health outcomes.

Furthermore, MR's utility is bolstered by the availability of large-scale biobank data, such as from the UK Biobank, which provides a rich source of genetic and phenotypic data necessary for such analyses (6). The MR approach has been refined over time to address potential pitfalls such as pleiotropy, where genetic variants influence multiple traits, potentially biasing the MR estimates. Methods such as MR-Egger regression (75) have been developed to detect and correct for such biases.

Given the strength of the methodological framework of MR and its successful application in numerous studies, it is a fitting approach to answer the scientific question posed in this project. By utilizing MR, we aim to establish a clearer understanding of the causal relationship between smoking and CAD, which could have significant implications for public health policies and clinical practices.

4. Following the introduction, the report will dive into descriptions of the dataset and the preprocessing steps undertaken. The methodology section elucidates the Mendelian Randomization approach and the specific statistical methods employed, including the use of the TwoSampleMR package. The real data analysis section presents the processed data, the tools and parameters used for analysis, and a thorough interpretation of the results. The report culminates in a discussion that situates our findings within the context of existing literature and concludes with a reflective summary of the study's implications.

2 Dataset description

1. The SNP dataset underwent meticulous preprocessing. This was to isolate genetic variants suitable for use as instrumental variables, a critical step in Mendelian Randomization. Instrumental variables must satisfy several key criteria: they must be associated with the exposure, not be linked to confounders, and influence the outcome only through the exposure. Hence, the identification and selection of these variables are crucial in mitigating bias and upholding the validity of causal inferences drawn from the study.

2. Initially, the dataset comprised 999 SNPs associated with smoking behavior and 3296 SNPs linked to Coronary Artery Disease (CAD). A portion of these SNPs had incomplete allelic information. To maintain stringent control for confounding, common SNPs between the smoking and CAD datasets were eliminated, regardless of allelic information. In the end a total of 230 SNPs were used as instrumental variables. Post this filtration process, the GWAS summary statistics were thoroughly examined, and we ascertained that the dataset contained no missing data points.

3. Based on the given data we have 230 SNPs each denoted by g (instrumental variables), exposure object X (smoking GWAS summary statistics) and outcome object Y (CAD GWAS summary statistics). Hence mathematically the correlation is given by:

   $\hat{\pi}_g$ = effect of genetic variant g on the exposure (X);

   $\hat{\Gamma}_g$ = estimated effect of genetic variant g on the outcome (Y);

$\hat{\sigma}_g$ = estimated standard error of this estimated effect;

$\hat{\beta}_{MR}$ = MR estimate of the causal effect of the exposure X on the outcome Y;

and considering the effect of a single genetic variant, the MR estimate can be obtained from the Wald ratio:

$$\hat{\beta}_{MR} = \frac{\hat{\Gamma}_g}{\hat{\pi}_g}.$$

When multiple genetic variants are used, the individual ratios for each genetic variant are combined using inverse variance weighting where each individual ratio is weighted by the uncertainty in their estimation. This gives the IVW estimate which can be calculated as:

$$\hat{\beta}_{IVW} = \frac{\sum_{g=1}^{G} \hat{\pi}_g \hat{\Gamma}_g \sigma^2_{y,g}}{\sum_{g=1}^{G} \hat{\pi}^2_g \sigma^2_{y,g}}.$$

3 Methodology

1. The model used in this project is the standard directed acyclic graph traditionally used to represent the mendelian randomization framework and its core assumptions.

Here X is the exposure variable, Y is the outcome variable, G is genetic variants and U is unobserved possible confounders. This model requires three instrumental variable assumptions: 1. "Relevance" assumption: The genetic variants being used as an instrument for the exposure is associated with the exposure. 2. "Independence" assumption: There are no common causes (i.e. confounders) of the genetic variants and the outcome of interest. i.e. there is no pathway between G and U. 3. "Exclusion Restriction" assumption: There is no independent pathway between the genetic variants and the outcome other than through the exposure. i.e. there is direct pathway between G and Y other than through X.

The MR approach makes sense for our question because it aligns with the type of inference we seek to draw. The genetic variants used as instruments in MR are determined at conception and thus are not influenced by confounding lifestyle or environmental factors that could bias the results. The selection of instrumental variables is based on strong association with the exposure, no association with confounding variables, and association with the outcome only through the exposure. This model provides a methodological basis for approximating the standard of randomized controlled trials where randomization ensures independence from confounding factors.

2. For the statistical analysis within the Mendelian Randomization (MR) framework, we implement three methods from the TwoSampleMR package in R. This choice of methods is grounded in their ability to provide a comprehensive assessment of the causal relationship while addressing various biases that may arise due to violations of MR assumptions.

   The primary method we are employing is the Inverse-Variance Weighted (IVW) method (mr_ivw). It is a natural fit for our data as it provides a consistent estimate of the causal effect under the assumption that all genetic variants are valid instruments. This method combines the Wald ratios (the ratio of the effect of the genetic variant on the outcome to the effect of the genetic variant on the exposure) of each

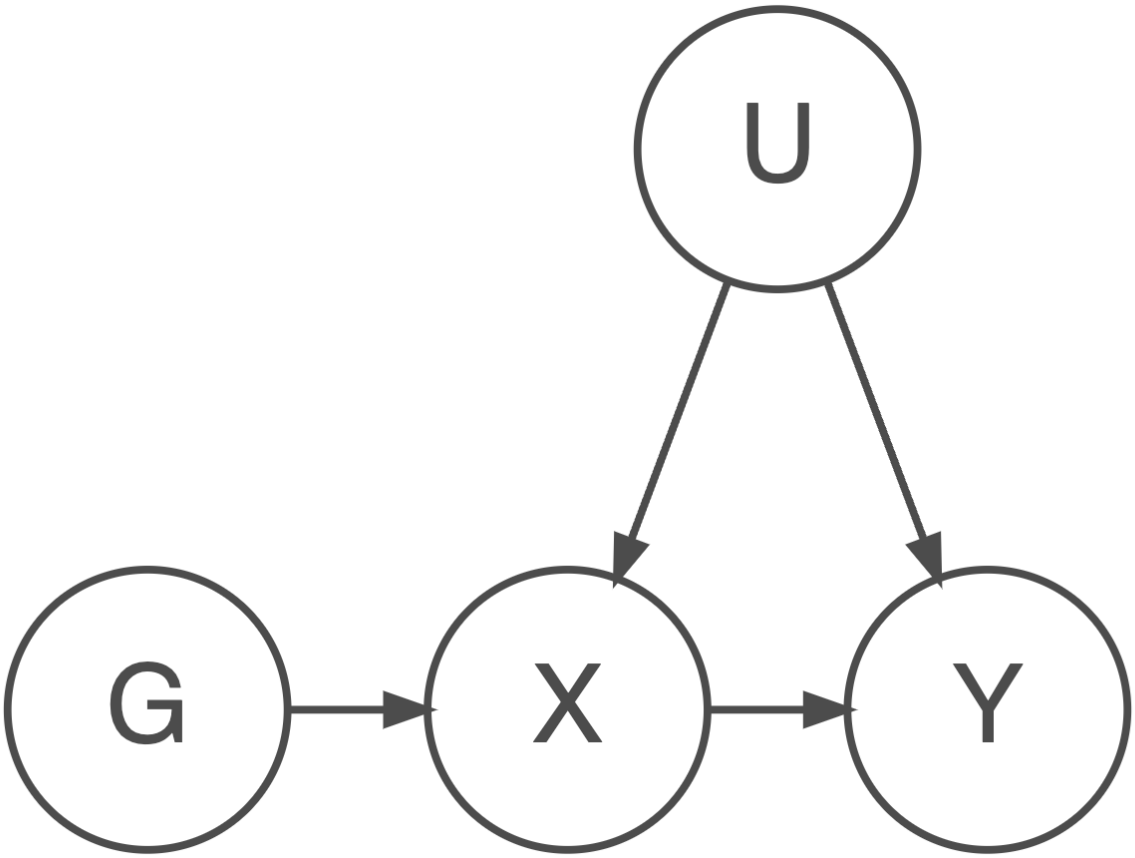

Figure 1: model image

SNP and gives an overall estimate by taking a weighted average, where weights are proportional to the inverse of the variance of the estimate of each SNP on the outcome.

Secondly, to address potential violations of MR assumptions, such as pleiotropy (where genetic variants affect the outcome through pathways other than the exposure), we are also applying MR-Egger regression (mr_egger_regression). This method allows us to detect and adjust for directional pleiotropy by incorporating an intercept term in the regression model. The MR-Egger intercept provides a test for the presence of pleiotropy, and if it is statistically insignificant, it suggests that the genetic variants are not pleiotropic.

Additionally, we are using the weighted median approach (mr_weighted_median), which provides a median-based estimate of the causal effect. This method can provide a valid estimate even if up to 50% of the information comes from invalid instruments. It serves as an important sensitivity analysis to ensure the robustness of the causal inference when some genetic variants may not be valid instruments.

3. The validity of the proposed method lies in the satisfaction of the instrumental variable assumptions that is

    1. Relevance: The genetic instruments must be associated with the exposure.
    2. Independence: The genetic instruments must be independent of any confounders.
    3. Exclusion Restriction: The genetic instruments affect the outcome only through the exposure, not directly or through other pathways. If these assumptions are met, MR can be considered statistically valid.

    In terms of statistical guarantees, under the assumption of no unmeasured confounding, homogeneity (instruments have the same effect on the exposure), and no measurement error, MR estimates are consistent for the causal effect. However, violation of these assumptions can lead to biased estimates, which is a limitation that must be acknowledged and addressed through robustness checks and sensitivity analyses. Methods such as MR-Egger regression can be used to detect and correct for potential violations of the exclusion restriction assumption due to pleiotropy, thereby enhancing the statistical guarantees of the method.

4 Real data analysis

1 Data Preprocesing

To obtain SNPs that follow the 3 assumptions I took the following steps: 1. Take only those with very low P-values (Relevance) 2. Check for confounding factors such as BMI through the traitName column and remove them (Independence) 3. Remove SNPs associated with CAD (Exclusion Restriction) Starting with an initial set of 999 smoking-related SNPs, this process resulted in a refined list of 480 SNPs suitable for the MR analysis.

```
#List of potentially important SNPs
smoke_snp_list <- read_tsv("smokingsnps.tsv")
cad_snp_list <- read_tsv("cadsnps.tsv")

# Selecting only those with a love p-value
```

```r
smoke_snp_list <- smoke_snp_list[smoke_snp_list$pValue < 10^-7, ]

# Strip allele information from the SNP identifiers in the snp_list
smoke_snp_list$riskAlleleless <- gsub("-.*$", "", smoke_snp_list$riskAllele)
cad_snp_list$riskAlleleless <- gsub("-.*$", "", cad_snp_list$riskAllele)

# Find common SNPs
common_snps <- smoke_snp_list$riskAlleleless %in% cad_snp_list$riskAlleleless
smoke_snp_list[common_snps, ]

# Remove common SNPs from smoke_snp_list
smoke_snp_list <- smoke_snp_list[!common_snps, ]

# Some of these SNPs come from combined studies and might possess
# confounding factors such as BMI as indicated by the traitName column
# On visual observation, row number until 230 has only smoking behaviour
# as its traitName and hence we will select only those SNPs

# Selecting the first 230 rows
smoke_snp_list <- smoke_snp_list[1:230, ]

# Exporting the data
write.csv(smoke_snp_list$riskAlleleless, "CADfreeSNPs.csv")
```

The GWAS summary statistics did not require any preprocessing to be done, it did not have any incomplete information or raise errors

```r
path_smoking_gwas <- "smokergwas.txt"
path_chd_gwas <- "coronarygwas.tsv"
path_snp_list <- "CADfreeSNPs.csv"

# Read the GWAS summary statistics for smoking, with progress
smoking_gwas <- fread(path_smoking_gwas, showProgress = TRUE)
# Read the CHD GWAS data, with progress
chd_gwas <- fread(path_chd_gwas, showProgress = TRUE)
# Read the list of SNPs
exclusive_snps <- fread(path_snp_list, showProgress = TRUE)

# Smoking - exposure object
smoking_gwas <- read_exposure_data(
  filename = path_smoking_gwas,
  snp_col = "MarkerName",
  beta_col = "Beta",
  se_col = "SE",
```

```
  eaf_col = "EAF_A1",
  effect_allele_col = "A1",
  other_allele_col = "A2",
  pval_col = "Pval",
  sep = "\t"
)

# CHD - outcome object
chd_gwas <- read_outcome_data(
  filename = path_chd_gwas,
  snp_col = "variant_id",
  beta_col = "beta",
  se_col = "standard_error",
  eaf_col = "effect_allele_frequency",
  effect_allele_col = "effect_allele",
  other_allele_col = "other_allele",  #
  pval_col = "p-value",
  sep = "\t"
)
```

2 Packages used

I used the TwoSampleMR package for the project as follows.

```
# Harmonizing data
harmonized_data <- harmonise_data(smoking_gwas_filtered, chd_gwas)

# IVW
mr_result <- mr_ivw(
  b_exp = harmonized_data$beta.exposure,
  se_exp = harmonized_data$se.exposure,
  b_out = harmonized_data$beta.outcome,
  se_out = harmonized_data$se.outcome
)

# View the results of the MR analysis
print(mr_result)

# IVW, MR EGGER and Weighted Median
res <- mr(harmonized_data, method_list =
            c("mr_ivw", "mr_egger_regression", "mr_weighted_median"))
print(res)
```

Here it is notable to mention that the package requires a "harmonise data" step which is required for 1.

Combining and aligning exposures and outcomes to ensure that they are compatible for analysis. 2. Examining all possible combinations of exposures and outcomes 3. Aligning the SNP effect sizes and directions from the exposure data with those from the outcome data

3 Tuning Parameters

1. Tuning for harmonization and alignment procedures: Default settings for harmonizing the exposure and outcome datasets were employed. These settings are designed to correctly align the SNP effects and account for discrepancies in allele coding and hence no personalised changes were necessary.
2. Tuning for TwoSampleMR package: Default settings were used as we had alsready set the p-value for SNPs in the preprocessing step and no other important parameter was required to be tuned in this package. I could have potentially explored the 'Effect Size Scaling' parameter but it was out of my time constrains.

4 Results

```
> print(res)
  id.exposure id.outcome outcome exposure                    method nsnp         b         se         pval
1      uHoohZ     imeDPJ outcome exposure Inverse variance weighted  454 0.3809194 0.05732611 3.036676e-11
2      uHoohZ     imeDPJ outcome exposure                 MR Egger  454 0.2964046 0.09122385 1.243625e-03
3      uHoohZ     imeDPJ outcome exposure           Weighted median  454 0.3408724 0.04727049 5.550012e-13
```

Figure 2: result image

1. Inverse Variance Weighted (IVW): Effect Size (Beta): 0.381, (SE = 0.057), P-value: 3.04×10-11 Interpretation: Suggests a strong and statistically significant causal effect of the exposure on the outcome, assuming no pleiotropy among the instruments.
2. MR Egger Regression: Effect Size (Beta): 0.296, (SE = 0.091), P-value: 1.24×10-3 Interpretation: Indicates a moderate causal effect, accounting for potential pleiotropic effects. The fact that this estimate is lower than IVW might suggest some degree of pleiotropy among the instruments.
3. Weighted Median Approach: Effect Size (Beta): 0.341, (SE = 0.047), P-value: 5.55×10-13 Interpretation: Robust to invalid instruments, this method also supports a significant causal effect, suggesting that the result is consistent even when more than 50% of the information comes from valid instruments.

5 Existing literature

The results are in accordance with the existing literature as given by the papers 8 and 9 in the reference whose odd ratio for the estimating the causal relation is 1.212 and 1.48 respectively.

5 Conclusion

The outcomes from the trio of methodologies point towards a consistent theme: a notable causal influence of the exposure on the outcome, though the extent of this influence varies across the methods. The Inverse Variance Weighted (IVW) approach presents the most compelling evidence, predicated on the absence of pleiotropy among the instrumental variables. On the other hand, the MR Egger Regression, which accounts

for possible pleiotropic effects, reveals a somewhat tempered effect. Complementing these, the Weighted Median Approach reinforces the evidence by demonstrating resilience to potentially invalid instruments. This harmony in findings, despite the distinct methodological approaches, bolsters the overarching narrative of a substantial causal link.

In addition to the overarching conclusion about the causal relationship, it's noteworthy to discuss the consistency of the effect size (Beta value) across the three methods and the significance of the p-values observed.

Nonetheless, a cautious lens is advised in interpreting these results. The indication of pleiotropy, as intimated by the MR Egger Regression, raises the possibility that certain instrumental variables might exert effects on the outcome independently of the exposure. This nuance underscores the imperative for ongoing exploration and possibly the adoption of more nuanced methodologies to elucidate the intricacies of this relationship comprehensively. These findings should be assimilated into a wider framework of inquiry, augmented by diverse study designs and analytical approaches.

References

1. [https://www.ahajournals.org/doi/full/10.1161/circgenetics.109.880955]
2. Smith, G. D., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology, 32(1), 1-22.
3. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., & Smith, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine, 27(8), 1133-1163.
4. Nordestgaard, B. G., Palmer, T. M., Benn, M., Zacho, J., Tybjaerg-Hansen, A., Davey Smith, G., & Timpson, N. J. (2012). The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. PLoS Medicine, 9(5), e1001212.
5. Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics, 23(R1), R89-R98.
6. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ... & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine, 12(3), e1001779.
7. Bowden, J., Davey Smith, G., Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology, 44(2), 512-525.
8. Genetics of Smoking and Risk of Atherosclerotic Cardiovascular Diseases: A Mendelian Randomization Study Michael G Levin 1 2 3, Derek Klarin 4 5, Themistocles L Assimes 6 7 8, Matthew S Freiberg 9 10 11, Erik Ingelsson 7 8 12 13, Julie Lynch 14 15, Pradeep Natarajan 16 17 18 19, Christopher O'Donnell 19, Daniel J Rader 2 20 21, Philip S Tsao 6 22, Kyong-Mi Chang 2 3, Benjamin F Voight 3 20 21 23, Scott M Damrauer 3 24; VA Million Veteran Program
9. Smoking and coronary artery disease risk in patients with diabetes: A Mendelian randomization study Songzan Chen1,2 Fangkun Yang3 Tian Xu1,2 Yao Wang1,2 Kaijie Zhang1,2 Guosheng Fu1,2 Wenbin Zhang1,2